Masters Program in Geospatial Technologies

INTEGRATING OPENSTREETMAP DATA AND SENTINEL-2 IMAGERY FOR CLASSIFYING AND MONITORING INFORMAL SETTLEMENTS

Brenda Ayo

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

INTEGRATING OPENSTREETMAP DATA AND SENTINEL-2 IMAGERY FOR CLASSIFYING AND MONITORING INFORMAL SETTLEMENTS

Dissertation supervised by

Joel Dinis Baptista Ferreira da Silva, PhD

Instituto Superior de Estatística e Gestão de Informação,

Universidade Nova de Lisboa

Lisbon, Portugal

Dissertation co-supervised by

Prof. Dr. Hanna Meyer,

Institute of Landscape Ecology,

Heisenbergstr. 2, D-48149 Münster

Dissertation co-supervised by

Ignacio Guerrero

Institute of New Imaging Technologies,

Universitat Jaume I

Castellón de la Plana, Spain

February 2020

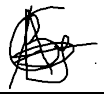
DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, Portugal, January 2020

Brenda Ayo



ACKNOWLEDGMENTS

Principally, I would like to thank my Supervisors, Joel Dinis Baptista Ferreira da Silva (PhD), Prof. Dr. Hanna Meyer and Ignacio Guerrero. You were instrumental from the inception, through the execution and end of this Thesis. I am extremely grateful and honored to have collaborated with you.

I am extremely grateful to all the professors that helped through this master's program especially professors Marco Painho and Christoph Brox for their guidance and support throughout the program duration.

Thank you to my cohort and fellow MSc. GeoTech graduate students – especially Damien and Chamodi - you have been incredible resources, whose positive and hardworking spirit has made my Master's experience so much more enjoyable and fun.

Lastly, thanks to my friends and family who were of great support even if you were thousands of miles away.

Integrating Openstreetmap Data and Sentinel-2 Imagery for Classifying and Monitoring Informal Settlements

ABSTRACT

The identification and monitoring of informal settlements in urban areas is an important step in developing and implementing pro-poor urban policies. Understanding when, where and who lives inside informal settlements is critical to efforts to improve their resilience. This study aims at integrating OSM data and sentinel-2 imagery for classifying and monitoring the growth of informal settlements methods to map informal areas in Kampala (Uganda) and Dar es Salaam (Tanzania) and to monitor their growth in Kampala. Three building feature characteristics of size, shape and Distance to nearest Neighbour were derived and used to cluster and classify informal areas using Hotspot Cluster analysis and ML approach on OSM buildings data. The resultant informal regions in Kampala were used with Sentinel-2 image tiles to investigate the spatio-temporal changes in informal areas using Convolutional Neural Networks (CNNs). Results from Optimized Hot Spot Analysis and Random Forest Classification show that Informal regions can be mapped based on building outline characteristics. An accuracy of 90.3% was achieved when an optimally trained CNN was executed on a test set of 2019 satellite image tiles. Predictions of informality from new datasets for the years 2016 and 2017 provided promising results on combining different open source geospatial datasets to identify, classify and monitor informal settlements.

KEYWORDS

Informal Settlements;

Remote Sensing;

Urbanization;

Machine Learning (ML);

Random Forest (RF);

Convolutional Neural Networks (CNN);

OpenStreetMap (OSM)

Sentinel-2 satellite imagery

ACRONYMS

CNN - Convolutional Neural Networks (CNNs)

DEM - Digital elevation model

DSM - Digital surface model

EO – Earth Observation

GEOBIA - Geographic object-based image analysis

GPS - Global positioning system

GSO - Global slum ontology

HR - High Resolution Satellite Imagery

ISDA - Informal settlement database atlas

ISUP - Informal settlement upgrading Programme

KOTAKU - Kota Tanpa Kumuh

LiDAR - Light detection and ranging

MGD - Millennium development goals

MKL - Multiple Kernel Learning

MLTs - Machine Learning Techniques

OBIA - Object-based image analysis

OHSA - Optimized Hotspot Analysis

OOA - Object-oriented analysis

OSM – Open Street Map

PSUP - Participatory slum upgrading Programme

RF - Random forest

RGB – Red Green Blue

SDI - Shack/Slum Dwellers International

SVM - Support Vector Machine

UAV - Unmanned Aerial Vehicle

UN – United Nations

UN-Habitat – United Nations Human Settlements Programme

VGG - Visual Geometry Group

VHR - Very high resolution

VHR – Very High Resolution Satellite Imagery

WDI - World Development Indicators

INDEX OF THE TEXT

INDEX OF TABLES

INDEX OF FIGURES

1. INTRODUCTION

1.1. Background and Motivation

Majority of the countries in the global south are the epicenter of global urbanization, which is without a doubt resulting into evident challenges which include high demand for employment, services and infrastructure [1, 2]. However, much of this urbanization is characterized by insufficient infrastructure, a lack of formal jobs, and haphazardly built and often squalid slums. Lack of planning, weak regulations, and, in some countries, the difficulty of obtaining title deeds for land, leads cities to grow out rather than up, making commutes longer and costlier which disconnects people and companies from jobs and markets, hence stifling the economy [3]. Most of these countries, especially in Africa are characterized by unbalance urban systems with one very large primary city and then less competitive smaller cities. The primary cities are predominantly crowded with people due to ease of access to resources and this has led to urban housing deficit accompanied by a huge infrastructure deficit. Informal settlements and slums develop because they are the only type of settlement affordable and accessible to the poor in cities, where competition for land is intense. [4] The UN-Habitat defines a slum as a group of UN-HABITAT defines a slum household as a group of individuals living under the same roof in an urban area who lack one or more of the following: (a) Durable housing of a permanent nature, (b) Sufficient living space, (c) Easy access to safe water in sufficient amounts, (d) Access to adequate sanitation and (e) Security of tenure that prevents forced evictions. Informal settlements and slums are usually characterized by inadequate basic Infrastructure and services such as water, sanitation, waste collection, storm drainage, street lighting, paved sidewalks and roads etc. [5, 6]. They also have inadequate access to schools, hospitals and public open spaces which are vital for human wellbeing [1].

Figure 1: Informal Settlements In Africa [7, 8]

Informal settlements / Slums are not a new phenomenon as they have been existent from the early years of urbanization and industrialization. The two main reasons for the growth of slums are Population growth and Governance [4, 6, 9].

Population Growth. As rapid urbanization is happening globally, many people move to the cities and urban regions in search for better facilities, employment and services [2, 4]. Rural – Urban and internal Migration occur due to peoples need for better job prospects, education, health facilities, or freedom from restrictive social or cultural realities. High birthrates in countries in the global south has lead most cities into an era of unprecedented growth [3].

Governance. Governments in Global South countries often cannot respond to rapid urbanization quickly enough, fail to recognize the rights of the urban poor and incorporate them into urban planning. This leads to the growth of Slums because the rate at which people are migrating to cities cannot be met by the planning process [4]. Some government and planning agencies also believe that providing services to the urban poor will only attract more urbanization and cause more slums to grow. However, this notion that people migrate to the cities for water or services is not accurate because most people are looking for employment [10].

Informal settlements are generally off the map in most cities because they are illegal and unplanned with undefined boundaries [6]. The growth of informal settlements is rapid and uneven over time and difficulties occur when the need to determine the number of those living in these settlements arises because most of them are not officially registered, which brings about uncertainties. This poses issues with regards to monitoring and planning of cities by the concern bodies which need adequate information in order to provide adequate services and infrastructure for those living in the urban areas as a whole. [11, 12] Informal settlements take on various shapes and sizes in different part world depending on various spatial characteristics of the areas in question but still possess some similar features and trends. These varying characteristics bring about complexities when creating unified methods that can be used across the board when dealing with different urban areas. These methods can be used to assess and predict the rapid urbanization by the different planning bodies and policy makers who are faced with inadequate information in different part of the world [11, 13]. In this Research, informal settlement and Slum will refer to the same thing.

To date, a significant amount of Research on informal settlement detection, monitoring and prediction has been carried out over the year with most recent studies based on urban modelling

using mathematical and machine learning algorithms. There has been a rise in the use of EO-based methods for mapping informal settlements, usually coupled with machine-learning methods for example [13, 16 ,17, 18 , 20, 21, 22]. They involved the use of EO data like Optical Space borne data, Synthetic Aperture Radar data, Unmanned Aerial Vehicle data etc., to distinguish of informal settlements or slums in urban areas by using either supervised or unsupervised machine learning image classification. Some object-oriented image analysis employed involved conjunction with local ontological expert knowledge. Other studies like in the development of the predictSLUMS model to identify and predict informal settlements uses street intersections data from Open Street Map, based on spatial statistics and a machine learning approach [20]. Other studies like Patel et al and A.T. Crooks [14,15] both developed geosimulation models that integrates a Geographic Information System and agent-based modeling (ABM) to examine the temporal changes of slums with developing countries cities in a particular spatial dimension and residential segregation respectively.

1.2. Research Gap Identification

As depicted from some of the studies shown above, several methods and techniques have been applied to map and classify informal settlements, and it's evident that solving the problem of identify and mapping informal settlements requires the collaboration of different parties like planning departments, GIS and Remote sensing techniques, Machine learning etc. The methods applied to detect and monitor informal settlements intend to distinguish these informal settlements from other urban areas based on concentration of housing units, road/ street networks. The location and mapping of informal settlements in most previous works mainly focuses on small portions of the City or Region with which prior knowledge of existence of the slum in that particular location is available. This poses a problem where there is no knowledge of the existence of an informal settlement in a particular area since the mapping is done partially. When more attention is given to already known slums, this allows for new slums to germinate since informal settlements tend to grow rapidly both in space and time, and yet there is already poor documentation on slums within the entire city. Existing strategies for mapping informal settlements oftenly require ancillary data such as location of schools, hospitals etc. in order to assess then based on remotely sensed data. However, this ancillary data is inadequate for poor developing countries. There is need for methods and techniques that can help to map the existence of informal settlements based on the distribution of geometrical and topological properties (Shape, Size, Orientation, Density) of their building

polygons, and at the same time identify informal regions across the entire city and not just particular places. There is also need to examine whether similar building polygons in informal settlements - similar in the sense of geometrical properties exhibit a spatial pattern. The informal settlement buildings characteristics coupled with Machine learning algorithms could help to predict where informal settlements could be and also if these characteristics are exhibited in varying geographical locations. This will help reduce the aggravation that comes up during their detection for inventory or planning issues and also provide more insight and understanding on the growth of these settlements.

The other challenge with mapping informal settlements at city, regional or country level is the absence of adequate data for analysis at these scales. The availability of open source satellite imagery like Landsat and Sentinel data provide a wider coverage but the spatial resolutions they offer does not provide for adequate distinguishing between formal and informal settlements in cities. However, with accurate delineating of these settlements at a finer scale, these satellite images can be used to monitor the growth of informality over time since the integration of spatio-temporal analysis of informal settlements is still inadequate.

1.3. Research Aim

The main aim of this research work is to integrate OSM data and sentinel-2 imagery for classifying and monitoring the growth of informal settlements.

To achieve the main research Aim, the following sub questions are addressed:

1. How can we exploit the potential of using buildings outline characteristics such as size and shape to differentiate Informal settlements from Formal Settlements?
2. Is it possible to predict informal areas in a city by understanding housing informality in other cities of similar context using buildings outline characteristics and machine learning?
3. What is the most appropriate Machine Learning technique based on accuracy to predict informal areas in a city based on buildings outline characteristics?
4. How can we exploit the potential of freely available Sentinel-2 satellite imagery with advanced machine learning to estimate the growth of Informal settlements?

1.4. Methodology overview

Based on the research questions, the following methodology workflow was adapted:

- Preprocessing of OSM data and Satellite Imagery. This involved data cleaning and preparation suitable for Analysis.
- OSM Building polygons enrichment with Feature geometric and topologic characteristics.
- Hotspot clustering analysis on building features training dataset to cluster Informal settlement buildings from the formal ones.
- Test and select the best classifier based on accuracy to predict informal areas in a city based on similarity context of buildings outline characteristics.
- Sentinel-2 image classification. This stage involves slicing the images into tiles and using the predicted Informal settlement regions from the step above to generate training set on image tiles of corresponding date period. Thereafter, a Convolutional Neural Networks Model is trained to identify Image tiles that belong to Informal settlements. The trained model is then used to predict Informal settlement image tiles for other temporal datasets.

1.5. Thesis Structure

The research is organized as follows:

- Chapter 2 reviews the related works and theoretical background on informal settlements, approaches for mapping and identifying informal settlements, and Machine Learning Algorithms
- Chapter 3 describes the study area, datasets and tools used for the research.
- Chapter 4 describes the implementation details of the methodology and the experiments conducted
- Chapter 5 presents the results;
- Chapter 6 addresses the analysis and discussion of the results including the limitations of the research;
- Chapter 7 presents a summary of the conclusions by answering the main re-search questions and recommendations for future works.

2. LITERATURE REVIEW

This chapter gives an overview on the fundamental concepts related to this research and the current state of research in the field of mapping informal settlements It highlights the physical characteristics of informal settlements, a discussion on various approaches used to map informal settlements and the integration of GIS and informal settlements and Machine Learning Algorithms.

2.1. Characteristics of Informal Settlements

The challenge with mapping and identify what an informal settlement is starts with the absence of a standard definition which has resulted into failure to incorporate these settlements into census and demographic surveys at national or even district level [23]. Based on literature, they take on names like Slum, Informal, Squatter, Spontaneous, Ghetto, Illegal, Irregular, or described by local names such as Favela, bidonville (French), mudun-safi (Arabic), bairros da lata (Portuguese) township or gecekondu (Turkish), to mention but a few [24]. The UN-Habitat qualitatively defined a slum as a household that is lacking either of improved water, improved sanitation, tenure security or overcrowded living environment is based on a measure of deprivation indicators. On the other hand, the definition of informal settlements differs from slum as those areas that developed through unauthorized occupation of land outside a legal, regulatory, planned and professional framework for example [26, 27]. Therefore, quantifying and classifying informal settlements/ slums in a universally acceptable way is particularly difficult because (a) What is referred to as a slum in one country may be a good quality of living and (b) over time and a settlement can get formalized through slum upgrading programs [27]. However, over the years, some features have been unique to informal settlements which include:

2.1.1. Building characteristics:

Buildings in informal settlements tend to be smaller (10 to 40 m^2), have simpler shapes that are mostly rectangular and heterogeneous orientation compared to those in formal settlements. The building density in slums is usually high since the spacing and gap between buildings in almost nonexistence [29, 30].

2.1.2. Access Characteristics:

Slums generally have irregular road access network with roads that vary in type, surface and width. [30]The roads are always short in length with many dead ends and dangles. The roads and access network in Informal settlements are usually narrow with limited usage by motor vehicles because they are either two narrow to let a car pass or can only allow access of on vehicle at a time. Slums are characterized with mostly footpaths.

2.1.3. Location and neighbourhood Characteristics:

Slums are usually located in hazardous areas like near dumping sites, wetlands, along railways, sewerage canals. [31]This is because most of the land where these features are located is public land and in abandoned. They are usually close to employment opportunities for unskilled and low skilled jobs like manufacturing industries to ease access to employment.

2.2. OpenStreetMap for Urban Analytics

Access to spatial data has changed rapidly over the years from traditionally prohibitive sources to openly licensed content and data due to changes in Information Technology and Communication (ITC) brought about by internet, social media and inexpensive portable GNSS devices like mobile phones and Handheld GPS [10, 32]. OSM is the most popular geospatial open data source platform containing billions of entries of VGI and is maintained by a massive community of mappers from around the world all working to-wards the goal of curating accurate and complete geospatial data. In the recent years, there has been an increased use of VGI data not only in GIScience but also in other fields like ecology, planning, computer science etc. [33,34,35]. for academic research and studies. OSM roads layers are associated with the highest completeness compared to buildings based on a recent study [36] that estimates that the OSM roads have reached more than 80% of completeness at a global scale. OSM datasets have been used to extract urban information and for analysis [37, 38]. These datasets are usually used as a supplementary source of information for land-use mapping with emphasis on classifying artificial surfaces [39], by combining OSM and satellite images to extract information related to the urban environment [39–42]. Few research studies exist on building footprints data enrichment for urban analysis [43, 44]. In these studies, automatic methods are developed by using the geometric and topological features in footprint data, in order to enhance the maps with the building usage information. Henn et al. [45] derived the architectural types

of buildings based on 3D coarse block models with vertical walls and flat roofs using Support Vector Machines (SVMs), whereby, geometric features such as length, width, area, and degree of perpendicularity of building footprints, types of buildings, as well as height information of buildings are required for the classification process.

2.3. Approaches for identifying and Mapping Informal Settlements

Studies on the development and existence of informal settlements have been carried out not only by geographers but also social scientists who put emphasis on social and economic aspects of informal settlements. Due to their complexity and ever varying nature, informal settlements are usually inadequately represented on urban and city maps which leaves them vulnerable to neglect by concerned government agencies. There are broadly three methods used to collect data to inform characterization and classification of informal settlements: household surveys and census, Participatory approaches and Earth Observation imagery analysis using geospatial Techniques [46].

2.3.1. Survey and census-based approach

These approaches use data collected through census and surveys like data on demographics, population, social and economic aspects to map deprivation and poverty. This data maybe coupled with data from official sites like the world Bank is used to map slums based on social, economic and habitat/Infrastructure variables. In [47], an exploratory factor analysis to define the Slum Severity Index (SSI) Mexico City based on measuring the shelter deprivation levels of households from 1990 to 2010. The results showed that the SSI decreased significantly between 1990 and 2000 as a result of several policy reforms but increased between 2000 and 2010. Weeks et al. [48] quantitatively used census data for Accra, Ghana to create a slum index based on the UN slum indicators to measure the concentration of slums. High correlations were found between the slum index, the socio-economic characteristics of neighborhoods and certain land cover metrics derived from VHR satellite imagery. A fusion of open sources physical and socio-economic data in [9] was used to develop an indicator database for characterizing slum settlements by leveraging data mining techniques for mapping slums in Kenya's major cities. It's important to note that one of the major challenges is related to census under-coverage due to time constraint, inadequate quality assurance and inaccurate addresses. The absence of

censuses in informal settlements is largely due to the inaccessibility of make-shift structures due to political intolerance or general disorder [30].

2.3.2. Participatory-based approach

The participatory approaches involve the cooperation and participation of the informal settlement dwellers in order to generate both spatial and non-spatial information to profile informal settlements [30]. The participatory slum upgrading Programme (PSUP), established by the UN-Habitat and the Slum Dwellers International (SDI) was a participatory approach that encouraged and empowered communities to become active partners with stakeholders in devising strategies to plan sustainable informal settlement upgrades [49]. The Kota Tanpa Kumuh (KOTAKU) platform and program is responsible for the handling of informality in Indonesia through increasing the role of local government and community participation. Informality is defined by buildings, local streets, drinking water supply, community drainage, wastewater management, garbage management, and fire protection [50]. However, Participatory-based approaches are extremely time consuming and effort-intensive is characterized by limited spatial coverage making it difficult to cover larger areas like regional or district levels [30].

2.3.3. GIS and Remote sensing-based approach

The connection between informal settlements and geography has been a focus of study for a long time. Many research studies have investigated the value of using Image- based identification of informal settlements through remote sensing and other GIS analysis tools for more than 2 decades [23,13,48,49, 50]. Three key criteria are often used in such studies: small grain size, high ground coverage and irregular access networks. Other factors such as proximity to hazardous areas, lack of vegetation and low- quality roofing are sometimes included. Such studies often conflate informality with slums. But while most such studies map a binary distinction between formal and informal morphologies, some adopt a more nuanced approach. Bradley et al [54] deploy the use of both HR and VHR satellite imagery to map informal settlements in developing countries using machine learning. They used the Canonical Correlation Forests (CCFs) to learn the spectral signal of informal settlements from HR satellite imagery. The second method used a CNN combined with VHR satellite imagery to extract finer grained features. Peter and Gulnaz studied the use of object based change detection and object tracking of informal settlements using remotely sensing data in Cape town (South

Africa) between the period 2000 to 2015. The object change detection provided for detection of thematic changes per object and to document the changes of their property values and potential movement. In [52], the authors discuss and evaluate various machine learning approaches and a combination of various features to detect slums. The classifications methods evaluated included multi-class and hierarchical to distinguish urban from other classes such as vegetation and water. [55] presented the use of CNNs for the detection of informal settlements from VHR i.e. 0.60m Quickbird satellite image for the city of Dar es Salaam, Tanzania acquired in 2007 and compared the results with state of the art classifiers that use handcrafted features. [56] evaluated and pointed out that Object Based Change Detection (OBCD) and OBIA are the most promising techniques for automatized identification of new buildings in informal settlements, but they need to work with VHR data rectified with a DSM of the same or compatible spatial resolution. [57] This paper aimed to document and understand basic infrastructural conditions and their changes over an eight-year period for the Kibera Slum using VHR satellite images by analyzing the dynamics of physical transformations of building sizes and heights, built-up densities, and building arrangements. A pioneering study by Hoffman [58] used OBIA to identify informal settlements from IKONOS imagery in the City of Cape Town. Informal settlement classification was undertaken using sub-classes that described settlement forms (dense, medium, new and bright) based on complex hierarchy and class descriptions such as textural and spectral features. The author found that the ability to detect informal settlements was dependent on the spatial resolution of the imagery. No quantitative results were presented as the findings of the study. This research was later improved by Hofmann et al. [59], who showed that several modifications were required when applying extraction methods to a Quickbird scene in Brazil. The adaptions included simplified and pruned class-hierarchies to make the chosen class descriptors in theory more transferable to comparable scenes. The results of this study demonstrated that the selection of a strategy for informal settlement segmentation and classification is data and context-specific. Shekhar [60] delineated informal settlements in Pune City, India, using Quickbird imagery. The study highlighted the efficacy of the developed methodology to discriminate informal regions by describing typical characteristics of these settlements. Fuzzy membership function of texture, geometry, and contextual information were used to achieve an overall accuracy of more than 87%. Kohli et al. [30] expanded upon the work of Hofmann et al. [59] and developed Generic Slum Ontology (GSO), which can be used as part of a conceptual classification OBIA schema.

However, techniques like OBIA require VHR Data usually with a resolution of under 5 meters to detect slum objects and is a very time consuming approach and requires an adoption of parameters and values to a AOI. This makes them difficult to transfer to other cities or countries. It should be noted that most of the recent approaches to identify and map informal settlements involve the fusion of two or more approaches.

2.4. Hot Spot Clustering Analysis

Hotspot analysis is a spatial analysis and mapping technique interested in the identification of clustering of spatial phenomena into *s*tatistically significant hot spots and cold spots [61].These spatial phenomena are depicted as points in a map and refer to locations of events or objects. A hotspot is an area that has higher concentration of events compared to the expected number given a random distribution of events [62]. Hotspot detection has evolved from the study of point distributions or spatial arrangements of points in a space [63] to understand spatial patterns in time. The application of hotspot analysis within public health, epidemiological research and crime mapping and research has increased significantly in the past couple of decades mainly due to advancement in GIS-based software. Hotspot analysis usually involves the incident count of points in a location, Attributes that further describes points or Period of time i.e. Date or time of events. Very few studies are based on the use of hotspot analysis in the mapping of Informal settlements because it requires mostly vector data which is quite insufficient for slums compared to raster data. Mohamed et. al introduced a method for identifying and predicting informal settlements using street intersections data using Getis-Ord Gi Hot Spot Analysis and a machine learning approach[20].

2.5. An overview of Machine Learning Techniques

Machine learning techniques (MLTs) (Figure 2) generate knowledge in a learning phase by means of training data and transfer this to new data for prediction and generalization even with noise-contaminated and incomplete data [58,59].

MLTs have been successfully used in analyzing and modeling various complex environmental disciplines, including in medicine, financial markets, ecology, geography, biomedicine, and epidemiology [66–69]. The growing popularity of MLTs can be attributed to their abilities to approximate almost any complex non-linear functional relationship [70–72]. Machine

Learning algorithms are based on both supervised and unsupervised learning. Supervised Learning works with labeled input data, known as training data in a training process for predicting unknown data. Based on the known data, the training process thereby continues until the model achieves a desired level of accuracy. Once the model is trained it can be used to transfer this knowledge on new data by generalizing from the training data to unknown examples. Supervised learning problems can be further grouped into classification and regression problems. On the other hand, for Unsupervised Learning, the input data is not labeled and does not have a known result. The model tries to recognize a pattern in the input data to learn about it and extract rules. This can be done through a mathematical process to systematically reduce redundancy or by organizing the data by similarity like through Clustering and Association.

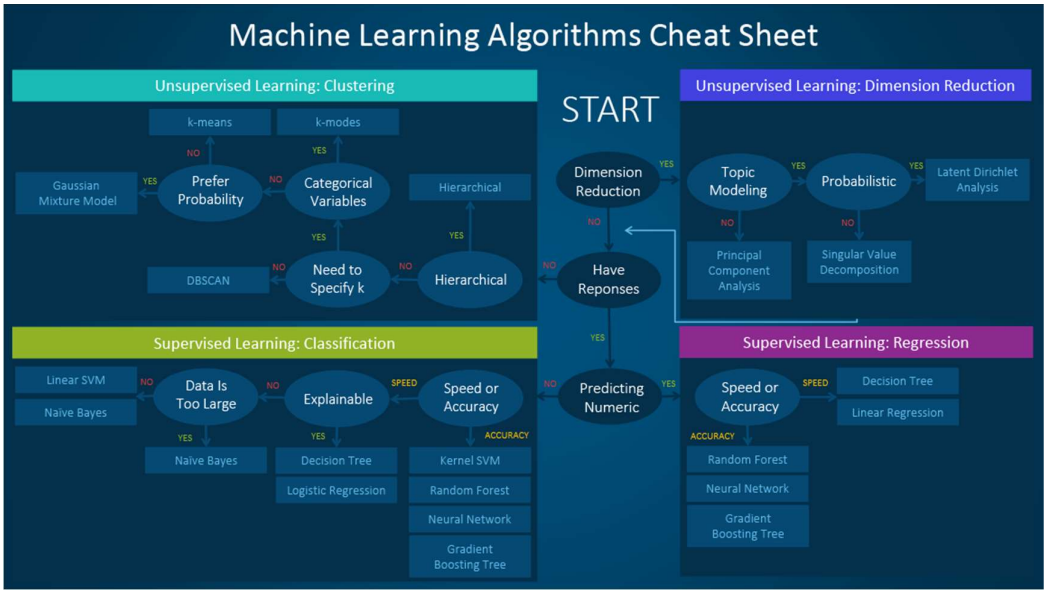


Figure 2: Machine Learning Cheat Sheet [73]

Some of the most common ML algorithms in Geospatial Sciences and Mapping of Informal settlements are Logistical Regression, Decision Trees (DT), self-organizing maps (SOM), random forests (RF), support vector machines (SVM) and artificial neural networks (ANN). It's important to note that selecting the best algorithm for a task is quite challenging since no algorithm outperforms others in a given task. Machine learning techniques have been and are still being used to carry out geospatial analysis in various fields for clustering, classification

and prediction various environmental and socioeconomic phenomena using different types of data.

2.5.1. Convolutional neural-networks

Convolutional neural-networks (CNNs) are a branch of artificial neural networks designed to recognize objects in images based on their ability to develop an internal representation of a two-dimensional image [74]. The convolutional neural network is composed mainly of three types of layers i.e. Convolutional Layers, Pooling Layers, Fully-Connected Layers. The Convolutional layers are comprised of filters which are the neurons of the layer and a feature map which is the output of one filter applied to the previous layer. The pooling layers down-sample the previous layers feature map by following a sequence of one or more convolutional layers with the intention to consolidate the features learned and expressed in them. They are used to reduce the size of the data while maintaining the most important features. Fully connected layers are the normal flat feed-forward neural network layer used at the end of the network after feature extraction and consolidation has been performed by the convolutional and pooling layers, to create final non-linear combinations of features and for making predictions by the network [75]. CNNs are well suited for the task of image classification in the field of remote sensing, However, CNNs also pose challenges when applied to remote sensing as it is often difficult or impractical to obtain a large set of labelled images, the model performs poorly and tends to over fit [74]. Such challenges can be curbed using techniques like Data augmentation using rotations and flips to increase the total number of training images to further avoid overfitting problem [33].

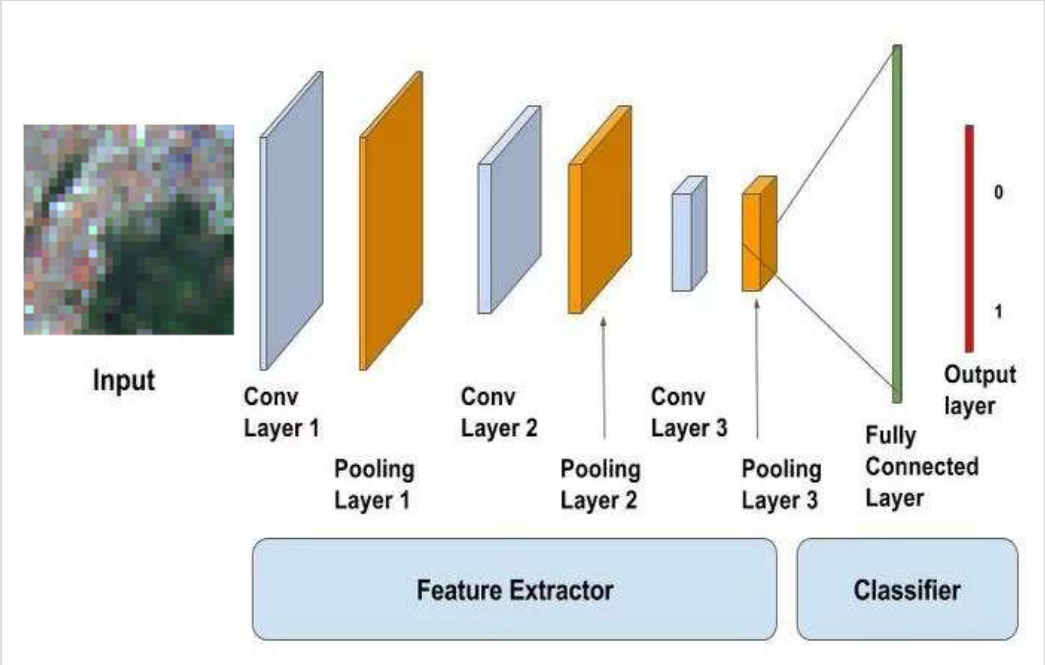


Figure 3: CNN Model Illustration [74]

Recent publications applying machine learning to remote sensing data, in particular to satellite imagery focusing on detecting or mapping informal settlements,[13, 55, 76–79] have typically been trained on a specific region, or feature in combination with VHR [80–82]. The approaches most in spirit to our own are [78, 79, 83]. Varshney et al. [78] focus on detecting roofs in Eastern Africa using a template matching algorithm and random forest, they take advantage of Google Earths' API to extract high resolution imagery, which although is free to researchers, is not openly available to everyone. Xie et al. and Jean et al.[19, 79] use a mixture of data sources and transfer learning across different data sets to generate poverty maps by taking advantage of night time imagery through the National Oceanic and Atmospheric Administration (NOAA) and daytime imagery through Google Earths' API. However, to our knowledge there exists no previous work on predicting informal settlements solely from LR data, or predicting informal settlements in the way that we present here. This inhibits our ability to benchmark against previous methods. Thus, by providing the data sets and the baselines in this paper, we provide a robust way to compare the effectiveness of any future approaches and facilitate the creation of new machine learning methodologies.

Informal settlements are still growing, however, the possibility for further improvements is rising with every new improved method. Therefore, this research attempts to reveal a new scope of mapping and identifying informal settlements by combining the benefits of Open data for mapping and monitoring.

3. DATA AND CASE STUDY

The following chapter presents the study area selected for research, the datasets and tools used This research is entirely conducted with open source geospatial data from OSM and Sentinel-2 satellite imagery.

3.1. Case Study

This research is mainly based on Kampala city, though the city of Dar es Salaam was selected as a second case study to assess the transferability of Building feature characteristics of informal settlements. The two East African cities of Kampala and Dar es Salaam were selected based on the availability and access to open source building polygons and their ranking and among the top fastest growing African cities. Kampala, the capital city of Uganda and Dar es Salaam (the major city of Tanzania) are African cities experiencing an era of unprecedented urban growth. Out of the world's 30 fastest growing cities, the top 10 are African including Dar es Salaam which is ranked second behind Kampala [3].

Kampala (Figure 4) with an estimated population of 1.5 million in 2014 and an area of about 8,451.9 km^2 [84] is divided into five divisions namely: Kampala Central Division, Kawempe Division, Rubaga Division, Makindye Division and Nakawa Division. The emergency of informal settlements in Kampala City has been gradual and sustained over a long period of time, which is attributed to the failure of Kampala Structure Plans to cater for the growth and development.

Dar es Salaam (Figure 4), the major port city of Tanzania is the center for industry, commerce and banking activities in the country. It has a population of about 4.36 million with an area coverage of about 1,590 km^2 [3]. It is made up of three districts Kinondoni in the north, Ilala in the Centre, Temeke in the south [85]. Dar es Salaam houses about 10% of the nation's population with approximately 70% of the urban households living in deprived areas [86,87]. More than three-quarters of residents live in informal settlements like Tandale.

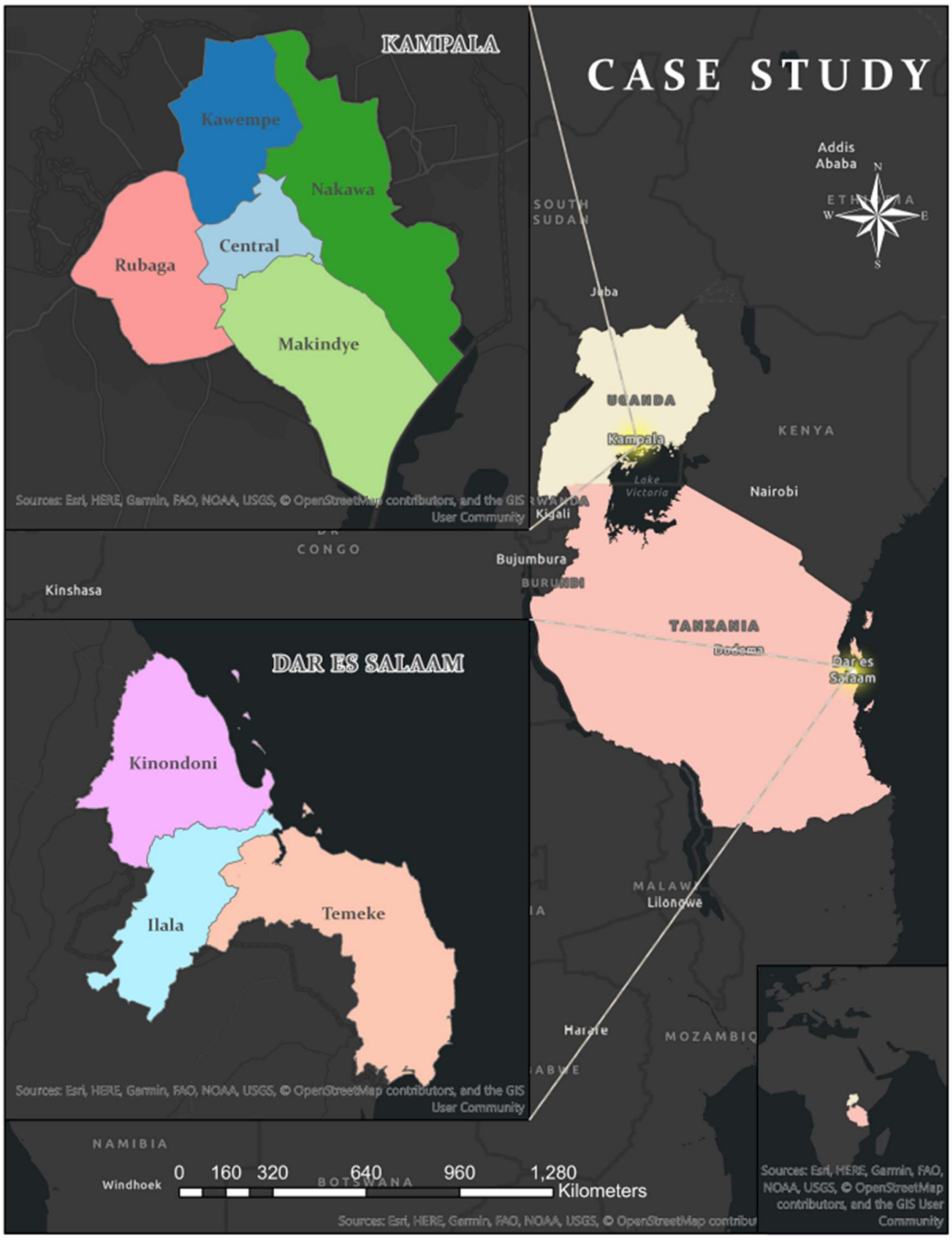


Figure 4: Case Study

3.2. Datasets and Pre-Processing

3.2.1. OSM Data

The datasets of Kampala and Dar-es Salaam cities were used for identification of informal settlement areas. For both cities detailed building outlines were downloaded from the Geofabrik website [88], which offers up-to-date countrywide shapefiles for geodata extracted from OSM. Detailed building polygons exist for the selected cities due to a recent release of building datasets for Uganda and Tanzania by Microsoft to the OSM database. Under the Microsoft's AI for Humanitarian Action program, a two-step process with semantic segmentation followed by polygonization resulted in 18M building footprints — 7M in Uganda and 11M in Tanzania. The building footprints generated from Bing imagery were detected using the Microsoft Cognitive Toolkit (CNTK) open source deep-learning toolkit and ResNet34 with RefineNet up-sampling layers [89,90]. Since OSM data is mostly captured by non-experts, some pre-processing steps were necessary in order to gain consistent datasets. Topological errors like Duplicate features and small gaps of buildings were fixed using both automatic and manual tools in ArcGIS pro and AutoCAD Map 3D [91]. The attribute tables were also examined to drop unwanted columns and also features that belonged to classes such as "*wall*", "*toilet*" etc. were dropped. A visual examination of the polygons was carried out to assess their correctness and completeness when overlaid on a basemap (Figure 5).

Figure 5: OSM buildings over ArcGIS Imagery Basemap

In Table 1, the total number of buildings for the datasets are listed. In order to create a training and test dataset for unsupervised classification, OSM data of Kampala was split into two partitions as shown in Figure 6.

City	# Buildings
Kampala (training partition)	214,885
Kampala (test partition)	214,191
Dar es Salaam	971,008

Table 1: Number of buildings and city per dataset

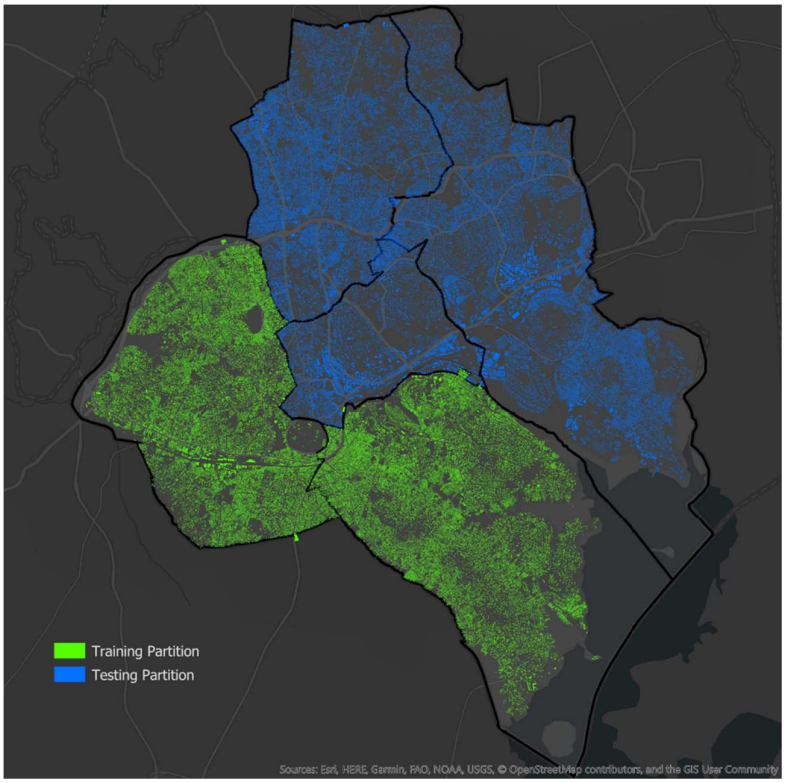


Figure 6: Partitions of the city of Kampala

3.2.2. Satellite Imagery

Several previous studies have successfully used Sentinel 2 data for slum detection and urban analytics [15,92] [93–95]. The pros of using sentinel-2 imagery lie in the fact that it offers a combination of spatial (10m), spectral (13 bands) and temporal resolution (5 days at the equator) based on two identical satellites [96]. Three (red, green and blue) of the Sentinel's 13 bands provide a 10m resolution, hence were selected for use in this research to monitor

Informal settlements, though most informal settlement buildings are smaller than this spatial resolution of 10m*10m. The images of the dates in Table 2 were selected based on perceived visually similarity, quality and minimal cloud cover.

Acquisition Date	Bands Used
01-01-2016	4,3,2
23-08-2017	4,3,2
31-12-2019	4,3,2

Table 2: Details of Sentinel-2 satellite images

3.2.3. Reference data

The map selected to validate regions mapped out as informal regions in Kampala (Figure 7) was prepared by AcTogether (Ug) in 2014, does not include the slum settlements which have recently come up in the city. However, it provides the locations of the slum settlements manually delineated through Fields surveys and Public Participation. The slums in Kampala city are not clustered at specific locations but have scattered presence all over the city with varying sizes and shapes. The settlements are located along the main roads in elongated and irregular shapes, along railway lines and near the lake. Unfortunately, recent Informality maps for Dar-es Salaam were difficult to come by.

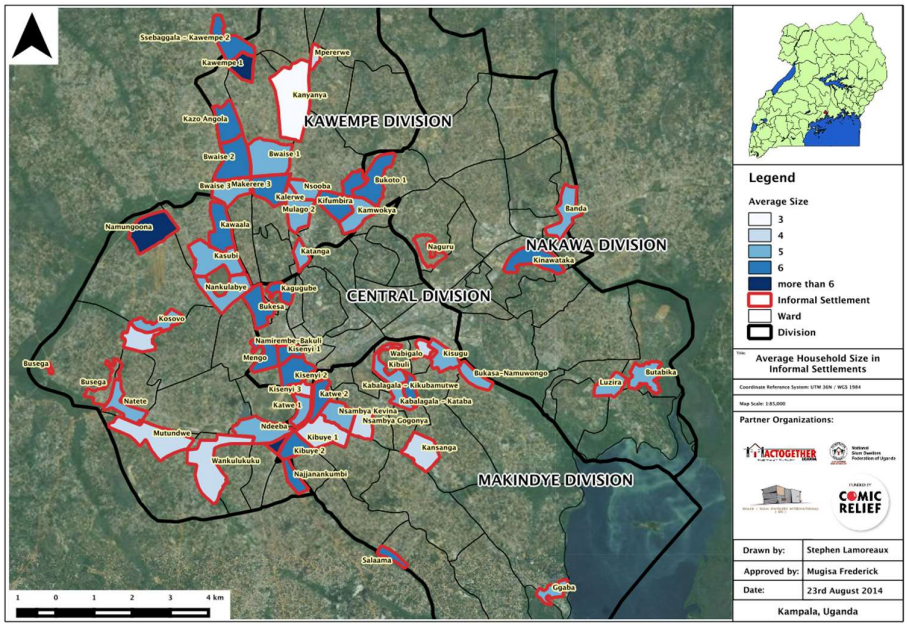


Figure 7: Distribution of Informal Settlements in Kampala based on Average Household Size [97]

4. RESEARCH METHODOLOGY

The method consists of three main phases (Figure 8). First phase consists of enriching the building outline features with geometric and topologic parameters,

The second phase entails hotspot clustering and classification of these characteristics to identify informal settlements in similar context cities, and the final phase involves spatio-temporal monitoring of informal settlements using sentinel-2 imagery.

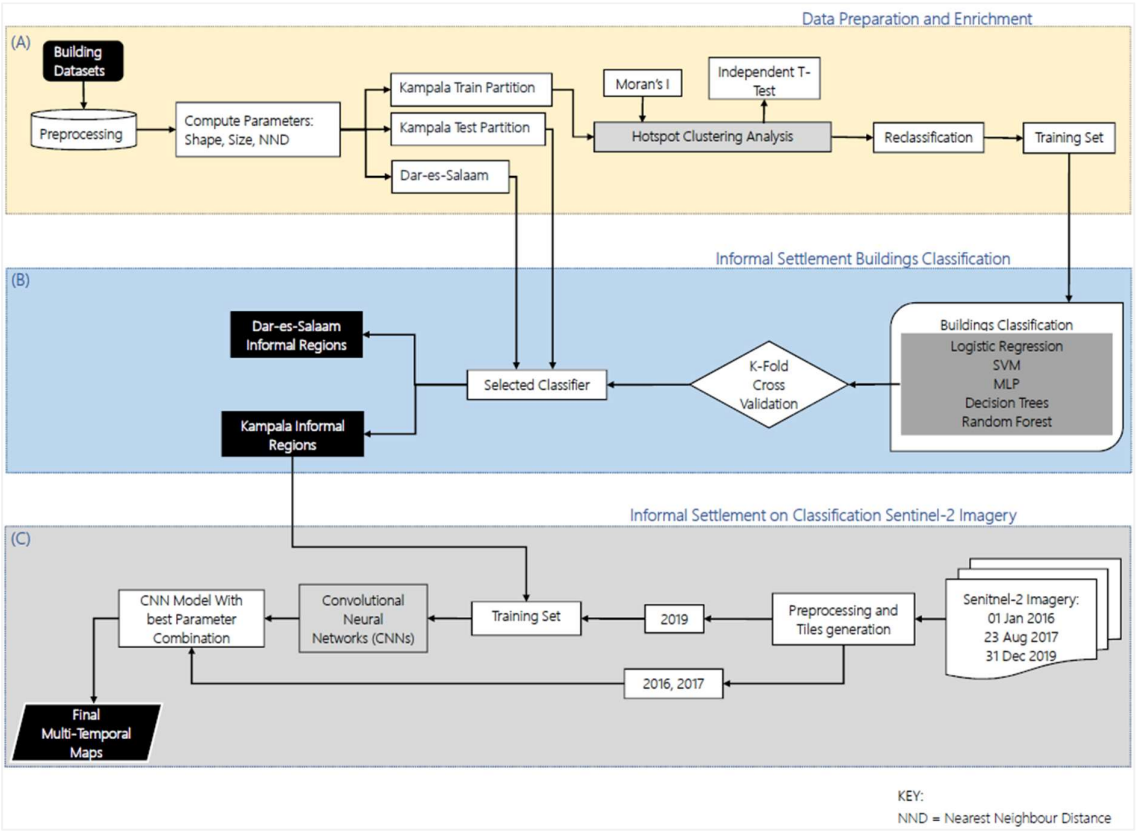


Figure 8: Methodology Flowchart

4.1. Data Enrichment and Computation of Parameters

Before carrying the clustering and classification process, geometric and topological measures to allow grouping of features into different clusters were determined. The measures chosen for enriching the building features were size, shape and shortest distance to a neighboring building. A building's shape was described by the number of vertices it has i.e. vertex number (VN). More complex buildings have many vertices and vis versa. The size was described by calculating its area (A) [98,99]. The Nearest Neighbor Distance (NND) and was used to describe the topological relations to neighboring features [100]. Buildings in informal

settlements tend to be very close to each other with limited gap between them. They are also small in size and have simple shapes like for rectangles or squares houses. The normalization of computed variables was carried out since they have varying scales and to avoid influence of some variables over others during use in the machine learning algorithms. The Min-max normalization approach was used where all the values were mapped between [0–1].

$$\boldsymbol{X' = \frac{(X - X\mathbf{min})}{(X\mathbf{max} - X\mathbf{min})}} \qquad \textbf{(}\text{Min-Max Normalization Equation}\textbf{)}$$

Where, X is an original value, X' is the normalized value, and X_{min}, X_{max} are the minimum and maximum values of this particular property.

4.2. Informal Settlement Buildings Clustering and Classification

In the first step, clusters are determined based on similarity of building feature characteristics using hotspot clustering analysis. The clusters are then reclassified to reduce the number of clusters. If a building belongs to the same cluster based on the features characteristics, its binned in one cluster and vis versa. This is done on the Kampala training partition. The second step involves using different classifiers to select one that suits the training partition data. The selected classifier is them used to determine informality regions in the Kampala Testing Partition and Dar-es-Salaam dataset.

4.2.1. Clustering of Informal Settlement Buildings

As shown in chapter 2, informal settlement areas have building features characterised by simple shapes, small buildings and have limited gaps between them.

In order to determine the pattern of spatial distribution and where low and high values of the building feature characteristics are grouped, hotspot analysis based on the use of Getis-Ord Gi* statistic was carried out on the Kampala training partition dataset to identify statistically significant hot and cold spots [61]. Since hotspot analysis requires the presence of clustering, there is need to test for the presence of clustering in the dataset by assessing spatial autocorrelation to identify clustering within the entire dataset. The Spatial Autocorrelation (Moran's) tool in ArcGIS measures spatial autocorrelation by simultaneously measuring feature locations and attribute values. If features that are close together have similar values,

then that is said to be clustering, and vis versa. The Moran's I returns values which include the z-score and p-value which will indicate if clustering is found in the data or not.

The Optimized Hotspot Analysis (OHSA) tool from ArcGIS Pro used for the hotspot analysis is identifies statistically significant hot and cold spots for multiple testing and spatial dependence using the using the False Discovery Rate (FDR) correction method [101]. The features are grouped into seven major clusters i.e. features in the +/- 3 Gi_Bins were statistically significant at the 99% confidence level; features with 0 for the Gi_Bin field was not statistically significant, those in the +/2 bins reflected a 95% confidence level, and features in the +/1 bins reflected a 90% confidence level. Upon computing these GiZscores where cold spots in all the feature characteristics most likely represent informal settlement buildings, a cross validation test using Independent samples t-test on the hotspot analysis results in order to differentiate between formal and informal zones and to determine whether there is statistical evidence that the associated group means are significantly different between them [102]. The GiZscores for both groups; informal, and formal regions are compared and when the p-value is less than 0.001, the null hypothesis can be rejected and the results of GiZscore can be illustrated as statistically significant [11].

The features identified as cold spots were then reclassified and grouped into one class and all the others as another class, which produces new reclassified results for the feature characteristics of Size, shape and NND. The three reclassified results are combined and if a feature has all its feature characteristics as cold spots, its grouped as informal and if not, its clustered as formal. This produces one resultant cluster dataset for the building features which shows both informality and formality in the Kampala training partition dataset. These results (especially the cold spots) were compared to areas delineated as slums in Kampala generated by a collaboration between Slum Dwellers International (SDI) and other organizations [46, 89].

4.2.2. Classification of Informal Settlements based on Building Characteristics

To predict where informal settlement regions are based on building features characteristics of shape, size and NND, based on the clusters from the training partition dataset, supervised machine learning classification was carried out on data of the same city (Kampala Testing Partition) and on data of a different city (Dar-es Salaam). Firstly, the clustering results were

applied to train, validate and test several classifier algorithms in order to select one that suits these datasets. Models were built using the python Scikit-learn library, taking advantage of the capabilities and features of this library. These models included Logistical Regression, Decision Tree (DT), Multilayer Perceptron (MLP), Random Forest (RF) and Support Vector Machine (SVM). The choice of these classifiers in based on their use in previous studies to classifier buildings based on their characteristics and also their ability to provide good accuracy with reference to chapter 2. The script involved splitting the Training partition dataset into 70% for training and 30% validation. Classification was performed first on the training data and accuracy measures, including overall accuracy, precision, recall and f-measure were noted. This was carried out in order to select the best classifier for the datasets. The selected classifier is then used to predict which features belong to Informal settlements in Kampala Testing Partition and Dar-es Salaam.

4.3. Monitoring the growth of Informal Settlements

The sentinel- 2 imagery after pre-processing and band composition were clipped to the Kampala Boundary. The acquired informal settlement regions from the classification phase above provides for polygons representing informality. We also assembled polygons for regions representing three categories: “built (without informal)”, “vegetation”, and “water”. All these regions were generated on the 2019 Image (Figure 9). This is based on the assumption that the OSM building outlines used to classify informality were as up-to-date as of October 2019. All the images were then sliced up into equal-sized tiles of 255*255 m to ensure uniformity in the datasets and placed in folders of *“Tiles_2019”*, *“Tiles_2017”* and *“Tiles_2016”* each with 5427 tiles. The tiles were saved as tiff files with attached geo-information which is required to plot the model predictions to the original locations in the map.

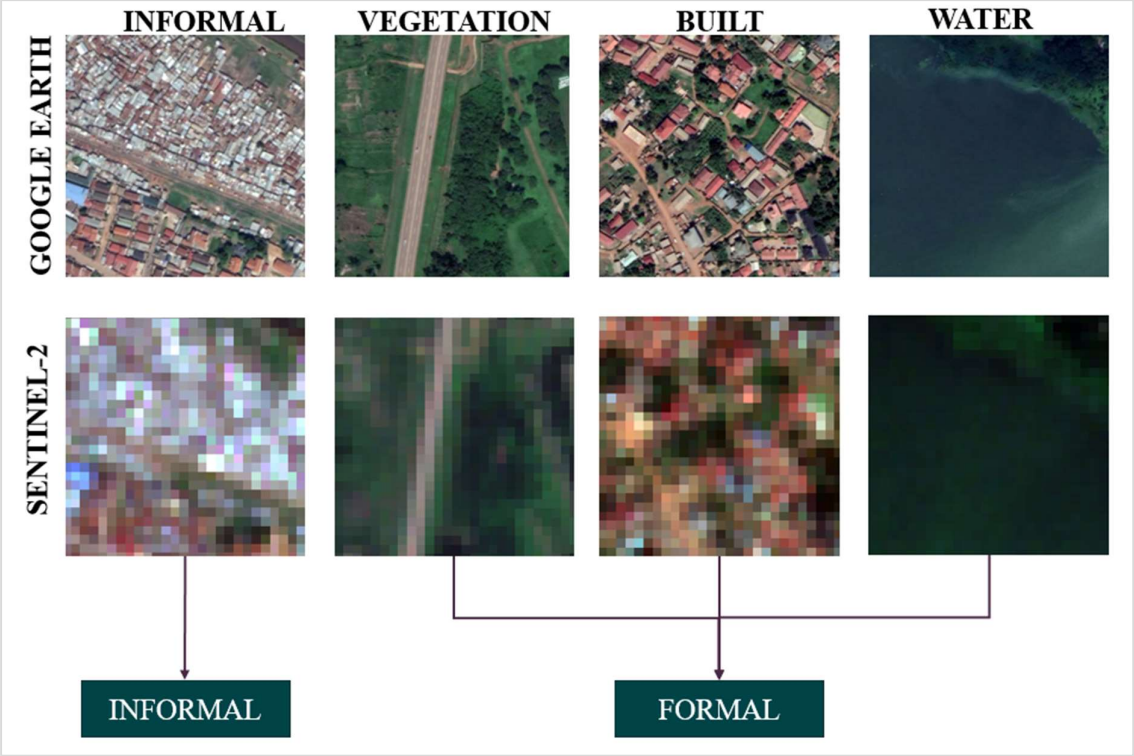


Figure 9: Informal and Formal Tiles

The tiles in folder *Tile_ 2019* were labelled based on the four categories created based on their intersection with the created polygons. All the tiles belonging to "built-up", "vegetation", and "water" were grouped into one folder as Formal and the Informal tiles placed in an Informal Folder. Regrouping the four categories into two was more accurate than training a binary classifier directly from the available data. The tiles were then split into training (80%) and testing sets (20%) for model-validation for use in the CNN model.

A baseline CNN model based on general architectural principles of the Visual Geometry Group (VGG) on which other models can be compared is established. The CNN architecture involved stacking convolutional layers with small 3×3 filters followed by a max pooling layer. These layers form a block, and these blocks can be repeated where the number of filters in each block is increased with the depth of the network such as 32, 64, 128, 256 for the first four blocks of the model. Padding is used on the convolutional layers to ensure the height and width shapes of the output feature maps matches the inputs. This function can then be customized to define different baseline models, e.g. versions of the model with 1, 2, or 3 VGG style blocks. The model was fitted with stochastic gradient descent, a conservative learning rate of 0.001 and a momentum of 0.9. Since the problem was a binary classification task i.e Formal and Informal,

the prediction of one value of either 0 or 1 was required. An output layer with 1 node and a sigmoid activation was used and the model optimized using the binary cross-entropy loss function. One-block (32 filters), two-block (32 and 64 filters) and three-block (32,64 and 128 filters) VGG models were all compared on the dataset. Two approaches to address overfitting of the training dataset: dropout regularization and data augmentation were explored to select the best suited one. Upon selecting the best model with the adequate model parameters, Predictions were performed on the Independent tile datasets for *"Tiles_2017"* and *"Tiles_2016"* to automatically predict informal settlement tiles out of all the tiles. A polygon layer was then built for the new informality tiles. This prediction step leads to tiles that are either formal or not. Polygons of the image tile Boundaries are created and those that belong to informal group are overlaid onto the map for visual interpretation.

5. RESULTS

This chapter describes the implementation details of the methodology and presents the results of this study. The source code of this implementation can be found at https://github.com/Bakked9/MastersThesis.

5.1. Clustering Analysis

5.1.1. Clustering Indices for Building Characteristics

The Moran's Indices for the Area, shape and NND are all close to 0.2, and have p-values of 0.0 (Table 3). The z-score is a standard deviation which measures how many SDs away an element is from the mean, while the p-value is a measure of probability that a random process created the observed pattern. A small p-value is an indicator that the spatial pattern is not random and vis versa. A positive Moran's I result indicates that neighboring areas or points are similar with respect to attribute values, while negative Moran's I shows that nearby regions are less similar in attributes than one would expect in a random pattern [103]. This goes to show that clustering exists within the dataset with respect to the measurement variables and hence Hotspot analysis can be carried out on the data.

Variable	Moran's Index	Expected Index	Variance	z-score	p-value
Nearest Distance	0.257656	-0.000017	0.000139	21.848243	0.000000
Area	0.278887	-0.000017	0.000139	23.696816	0.000000
Shape	1.336825	-0.000017	0.000139	113.285603	0.000000

Table 3: Spatial Autocorrelation (Moran's I) Results

5.1.2. Optimized Hotspot Analysis

The optimized hotspot analysis resulted into Gi_Bin fields which identified the statistically significant hotspots, along with the non-significant and cold spots, indicating the type of clusters for the feature characteristics. In the case of the Training partition, the Gi_Bin field showed that patches spread around the selected area was characterized by cold spots, non-significant types, and hotspots types that are distributed allover (Figure 10). The areas identified as non-significant cluster types provide some sort of boundary between the cold and

hot spots. Whereas the cold spots were located in a small area mainly on the north and central parts, the hotspot clusters were spread out in on the north, south, east and west sides.

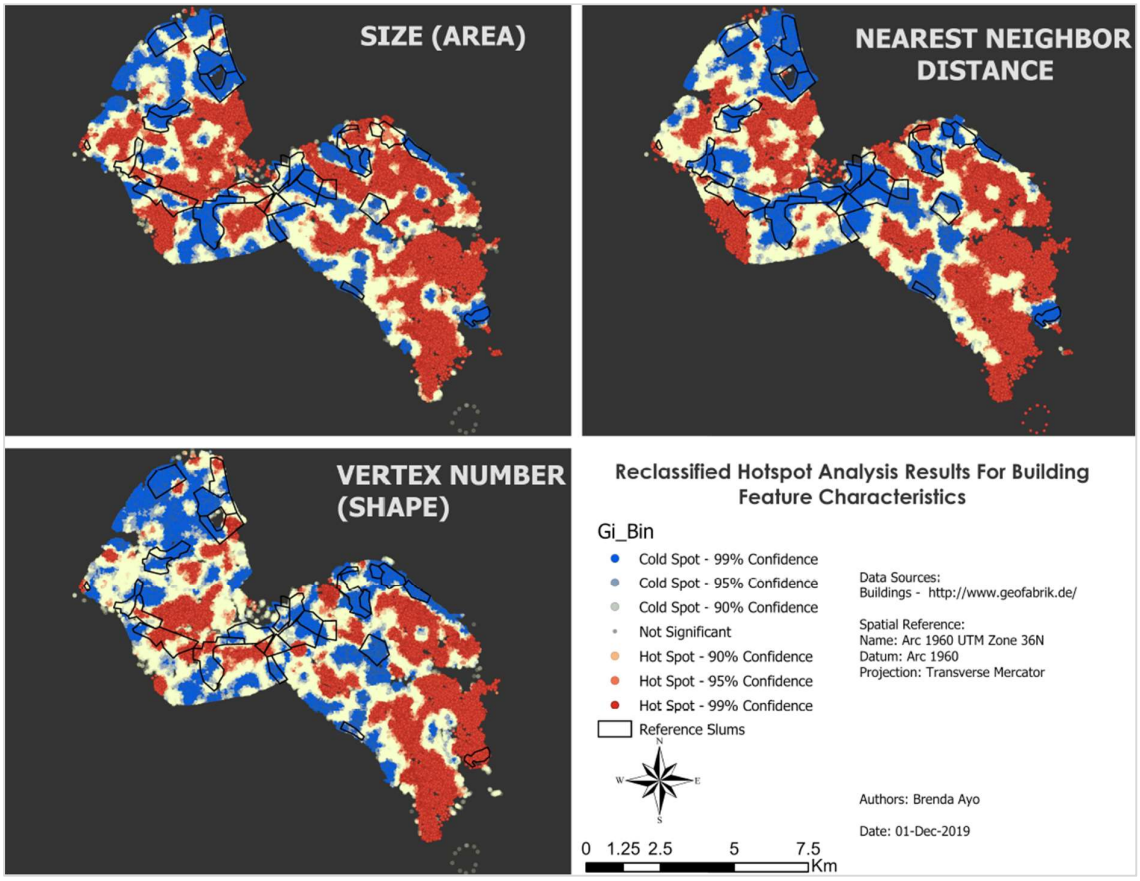


Figure 10: OHSA Results Maps for Size, Shape and NND

5.1.3. The Association between the Cold spots and Informal settlements

Table 4 illustrates the results of the t-test analysis, using a 95% confidence level for each feature characteristic. Since the Sig (2-Tailed) values i.e. the p-value for the GiZscore and Nneighbours predictors are very small i.e. $p<0.001$, we reject the null hypothesis of Levene's test and conclude that the variances in the GiZscore and Nneighbours predictors of Formal is significantly different from that of Informal. The Informal groups mean numbers of neighbors are larger than those of the formal areas. The mean GiZscores are the informal areas are also smaller than those of the formal areas. This shows that the informal regions are more likely to be cold spots with low values of Shape, size and NND.

		SIZE(AREA)				
Group		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**	**Sig.(2-tailed)**
GiZScore	ColdSpot(Informal)	104884	-4.4855	2.0227	0.0062	0.000
	HotSpot (Formal)	36868	3.6859	1.8595	0.0097	
Nneighbors	ColdSpot(Informal)	104884	958.00	302.407	.934	0.000
	HotSpot (Formal)	36868	363.13	129.434	.674	
		NEAREST NEIGHBOUR DISTANCE (NND)				
Group		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**	**Sig.(2-tailed)**
GiZScore	ColdSpot(Informal)	123467	-5.6911	2.8813	0.0082	0.000
	HotSpot (Formal)	41942	5.1848	3.3074	0.0161	
Nneighbors	ColdSpot(Informal)	123467	936.37	281.848	.802	0.000
	HotSpot (Formal)	41942	344.97	119.569	.584	
		SHAPE (VERTEX NUMBER)				
Group		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**	**Sig.(2-tailed)**
GiZScore	ColdSpot(Informal)	129377	-4.8803	1.9626	0.0055	0.000
	HotSpot (Formal)	34552	5.8729	4.6398	0.0250	
Nneighbors	ColdSpot(Informal)	129377	850.26	319.712	.889	0.000
	HotSpot (Formal)	34552	512.44	286.337	1.540	

Table 4: Independent samples t-test analysis Results

Cold spots in this analysis represent areas that have low values of Area, Shape and NND. Informal settlements are characterised by buildings that have simple shape, small is size and are densely located in a particular location i.e. short NNDs between polygons (see figure 10). With this, the cold spots resulting from the OHSA are considered to be locations with informality based one the building typologies used to characterise them.

To simplify the data for further analysis, all the OHSA results were classified to bin the data points into two classes i.e. “Informal” and “Formal”. This was done through reclassification of the Gi_bin results with a threshold of -2 or lower (95% cold spot confidence level) as Informality and the rest as formality. As shown in figure 11 and 12, the regions that identified as Informal share common locations across the map with respect to the building typologies. The results were combined to highlight only those areas that had been classified as informality in all the typologies. The resultant dataset was then used as a training set for the Classification step.

Figure 11: Buildings Classified as Formal and Informal

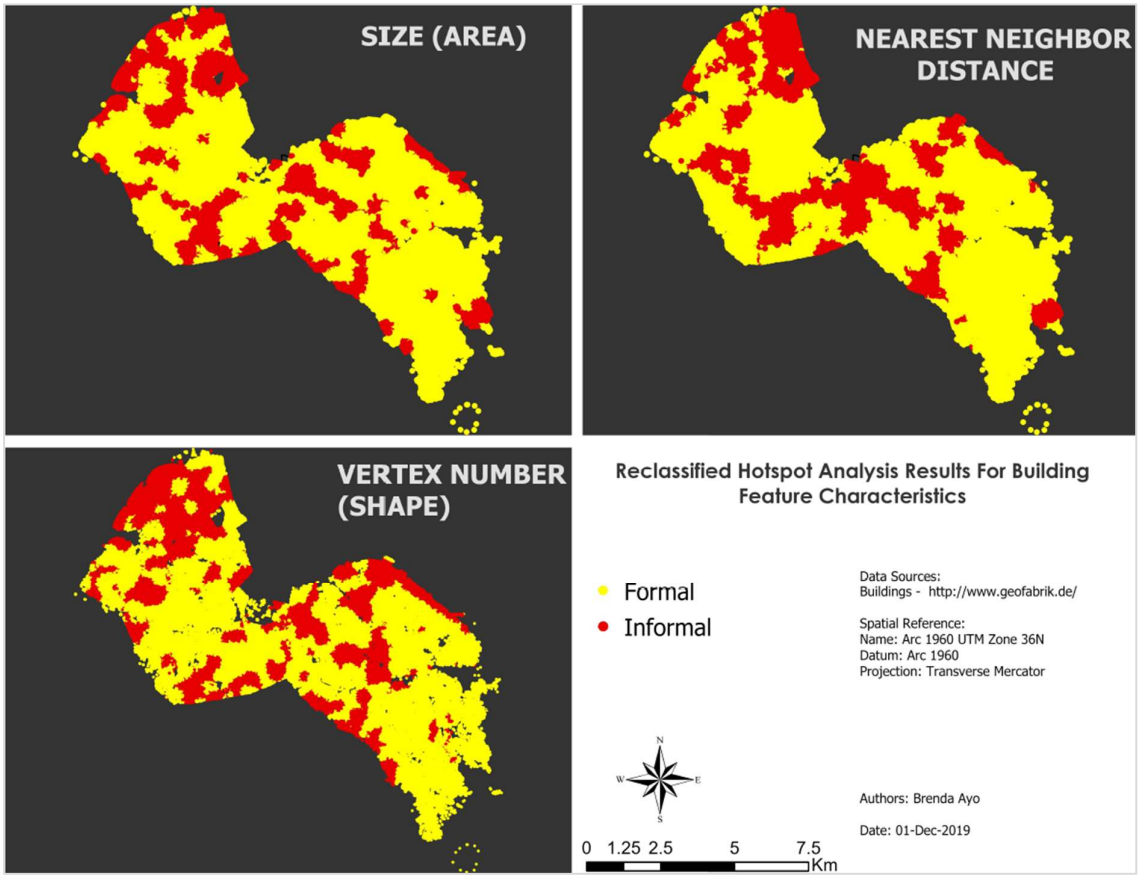


Figure 12: Informal Areas based on the Building Property Variables

5.2. Prediction of Informal Settlements using Machine Learning

The training data set was split into train and test to in order to enable training and validation of the models. Table 5 shows the accuracies obtained by training the data with all the algorithms. The highest accuracy was obtained using the Random forest classifier i.e. 89% and the lowest obtained from Logistical regression (74%) and Support Vector Machine (75%). In order to assess the skill of the model on new data, the K-Fold Cross validation was carried out and the Test-split from the original Training dataset (Figure 13). To further assess the performance of the model, a confusion matrix has been computed for the actual and predicted values for each category (formal/informal) for the entire dataset using the pre-trained models for each city, whereas an overall validation accuracy of the entire dataset is also illustrated.

Classifier	Accuracy	Precision	Recall	F1-Score
Logistical Regression	0.74	0.54	0.74	0.62
Decision Tree	0.84	0.84	0.84	0.79
Random Forest	0.89	0.88	0.89	0.88
ANN(Multilayer Perceptron)	0.81	0.79	0.80	0.77
Support Vector Machine	0.75	0.56	0.75	0.64

Table 5: Accuracy Assessment of Classifiers

Classifier	Logistical Regression (logreg)	Decision Tree (tree)	SVM	Random Forest	ANN
Mean accuracy	0.7484	0.8469	0.7484	0.8897	0.7484
Variance	0.0008	0.0193	0.0008	0.0163	0.0168

Table: K-Fold Cross Validation on the Classifiers

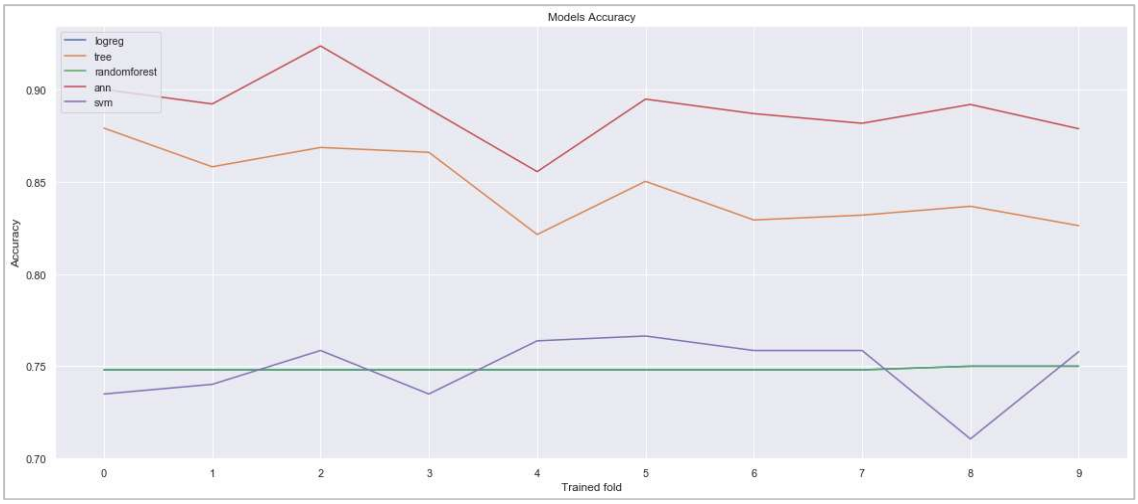


Figure 13: K-Fold Cross Validation of Classifiers

Based on the accuracy assessment results above, the Random Forest Classifier performed better than the other classifiers and hence was used to classify the data in the Kampala testing partition and the city of Dar es Salaam. The prediction accuracies for the two cities (table 6) were good enough to comfortably continue with Analysis

Town	Accuracy	Precision	Recall	F1-Score
Kampala_Test Set	0.81	0.80	0.81	0.78
Dar es Salaam	0.79	0.78	0.79	0.78

Table 6: Accuracy Assessment on Testing Sets

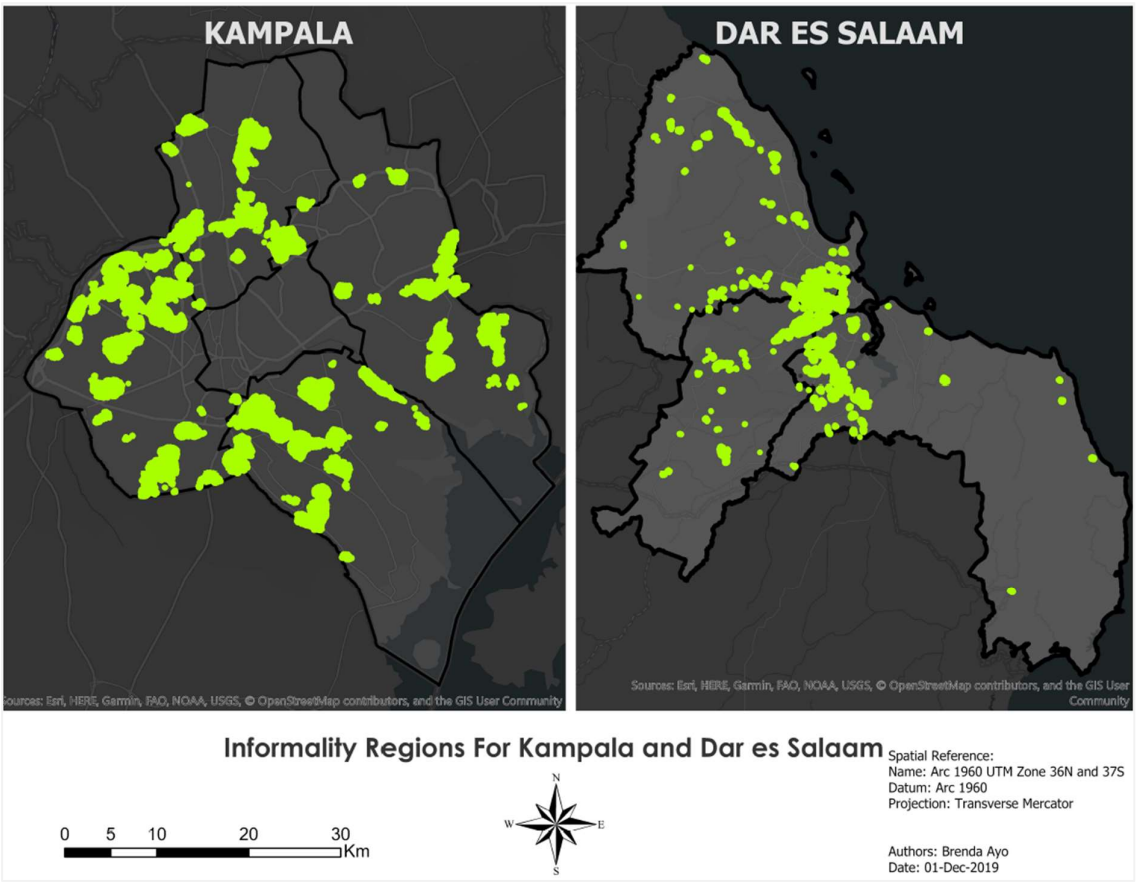


Figure 14: Informal Regions in Kampala (Left) and Dar es Salaam (Right)

The Informal settlement regions in both cities of Kampala and Dar-es-Salaam is shown in figure 14. The informality regions in Kampala are in all four Divisions except the central region. The central region is mainly characterized by already planned areas with planned streets, and mostly elevated buildings from the colonial times. It is also the Central Business District (CBD) where Business and Administration of the city is carried out. This little or no idle land on which informality may grow. For Dar es Salaam Informality is mostly shown in the center of the city across all municipalities. The largest Slum Tandale (located in the south of Kinondoni) was highlighted by the Classifier Model which shows that it is a good enough model.

5.3. Monitoring Informal Settlement Growth

The image tiles representing informal settlements were mostly composed of compact buildings and tiles appear compact with a mixture of brown and light pixels (Figure 15). Tiles comprising of relative dense areas of built-up > 50 percent and vegetation < 30 percent were

identified as those overlapping with the Informal regions hence classifying them as Informal areas. False positives existed in the study area that also comprising of areas of dense buildings and bare-soil areas due to similarity in pixel intensity. The image tiles used to train the CNN models as shown in figure 15 show a significant difference between formal and informal regions.

CNN Model	Accuracy (%)	Precision	Recall	f1-score
1-VGG	84.547	0.704	0.70	0.676
2-VGG	84.327	0.6395	0.74	0.6414
3-VGG	86.534	0.7271	0.77	0.6085
Baseline VGG3 + Dropout	80.353	0.774	0.73	0.6965
Baseline VGG3 + Data Augmentation	90.3	0.8214	0.79	0.7545
Pre-Trained VGG16	68.433	0.5915	0.64	0.5117

Table 7:Accuracy Assessment of CNN Models

The training dataset generated was used as input for training different CNN models after randomly split of 80% (training) and 30% (testing). The models were fine-tuned implementing varying parameters to improve the model. Table 7 summarizes the overall accuracy for each model with the use of a pre-trained VGG16 model achieving the lowest accuracy with 68.433% the second lowest was Baseline VGG3 + Dropout model with 80.353%. The best accuracy was achieved with the 3-block VGG model combined with Data Augmentation i.e. 90.3%. The three VGG-based architectures offer improved performance with increase in capacity. Since the training dataset was small, artificially expanding its size by creating modified versions of images (Image Data Augmentation) improved the accuracy of the model. The predictions made on the new images for the years 2016 and 2017 produced promising results for predicted tiles as Formal and Informal (Annex 8.3). Tiles that were predicted with a probability of $\geq 70\%$ as informal were all considered as belonging to informal regions. This was done to avoid inclusion of a lot of tiles with mixed categories especially because most boundaries between the categories are fuzzy in nature.

Annex 8.3 show the image tiles predicted as Informal settlements, represented as polygons. With visual inspection, the quantity of polygons increased through the four-year period. Informal settlements grew outwards to the neighbouring image tiles over the years. The North-

eastern and central parts of the study area have the least concentration and change in informal settlement regions. Since the central region is the CBD of the city, it also experiences limited development of Informal settlements over the period of time.

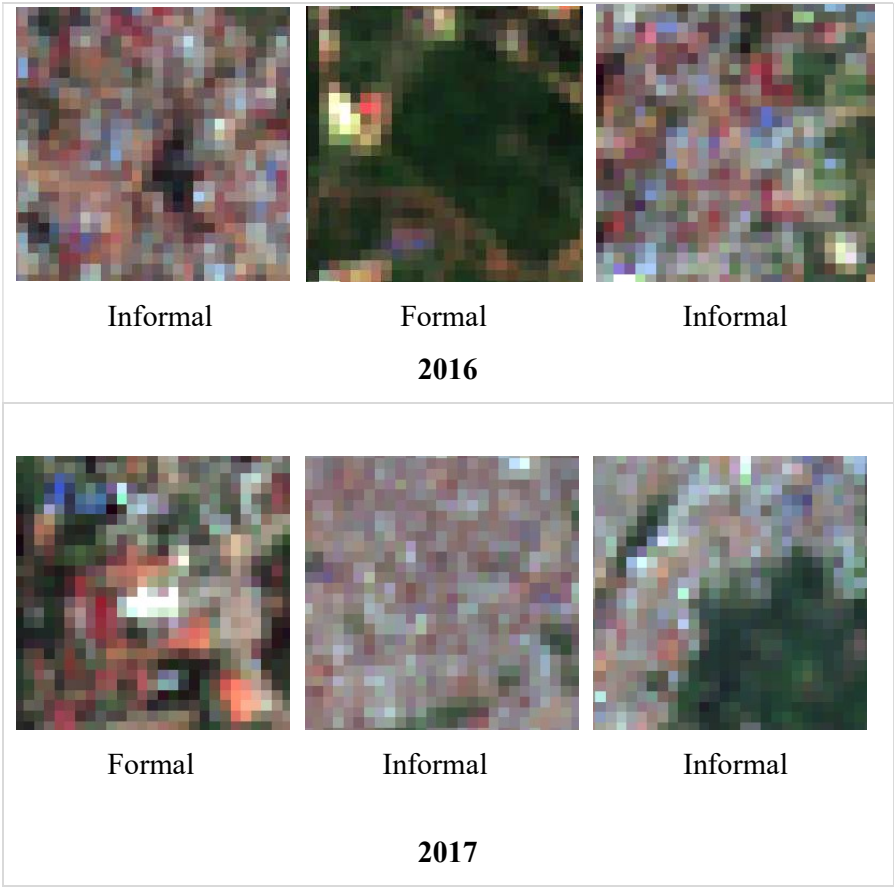


Figure 15: Classified Image Tiles

Figures 16 shows the changes of informal regions from 2016 to 2019. The maroon squares represent informal regions in January 2016, whilst the pink squares represent informal regions that developed by August 2017 and the cream squares represent informal regions that developed by the end of December 2019. The expansion of informal settlement regions is evident in areas Namuwongo and Bwaise. The map shows that the relative location of informal regions has not changed over time even if they now occupy more land. Informal regions remain on the fringes of the city.

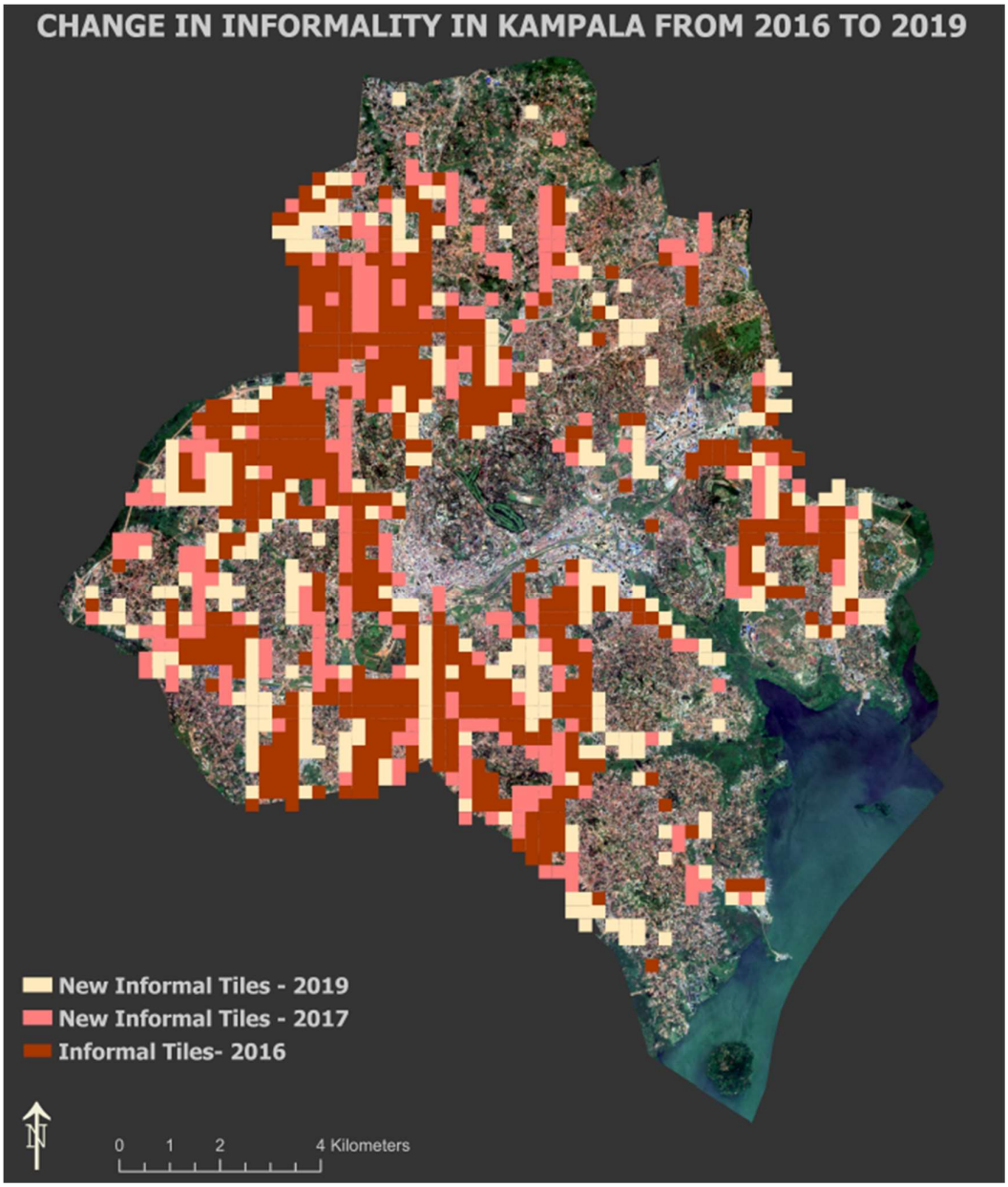


Figure 16: Change in Informal Settlement Regions in Kampala. This shows Informal tiles that existed in January 2016, new informal tiles that had developed by August 2017 and also by December 2019.

6. DISCUSSION

6.1. Discussion

This study presented approaches to classify informality regions based on building outline characteristics and monitoring growth of informal settlements over time by executing CNNs on High resolution satellite imagery. Firstly, we highlight that OSM building feature characteristics i.e. size and shape of building outline, and distance to nearest building can be used to differentiate between formal and informal buildings. In the approach, the geometric and topologic measures presented in this research were clustered based on the values high or low. As an important aspect of the evaluation, the transferability of the obtained clusters from training set to test data was investigated. Firstly, the results were applied to the testing data of the same city and then to another city. 108,390 of 429,270 buildings (25.25%) were classified as informal settlement buildings in Kampala while 188,728 of 971,008 buildings (19.44%) in Dar-es Salaam were classified as informal settlement buildings (Figure 13). There was a decrease in Random Forest classifier model performance on the Kampala Test Partition and Dar-es Salaam may be due to differences in the characteristics of buildings in very urban areas, or the fact that more data were available in Dar-es -Salaam. It's important to note that the accuracy values were consistently above 0.75, suggesting very good discriminatory performance irrespective of the city. The variation in building typologies used during this analysis (shape, size and spacing) played an important role during the classification of informality and formality even though these characteristics vary according to geographical location [93]. Though the morphological nature of informal settlements varies across the global, they still bare some similarities in characteristics. This could be the reason why the same model could be applied in a different city and produce distinct results. Compared to its areal coverage, there are fewer informal settlement buildings in Dar es Salaam. Over the years, a decrease in the population living in Slums in Tanzania has been decreasing with a sharp decrease observed from 2011 even with rapid population growth over time (Figure 17). This is probably due to strategies by the Tanzania government to shift from regularly demolished homes in informal settlements to awarding titles could be one of the reasons for a decrease in informality in the city. The MKURABITA Programme first launched in Dar es Salaam, aims at transforming property and businesses held in the informal sector into legal entities that are rooted firmly in the formal sector [104].

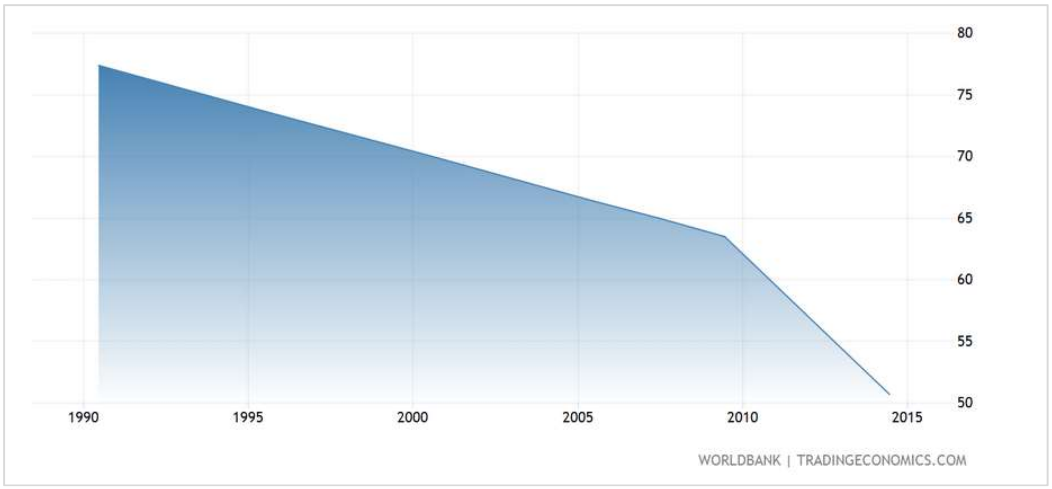


Figure 17: Tanzania - Population Living In Slums (% Of Urban Population) [[105]**]**

The experimentations of Using CNN on Sentinel-2 evaluates shows promising results on discriminating Informal from formal settlements over time. However, Seasonal variations between the images prevented the deep learning model from recognizing certain Informal Settlements. Testing the model on new multi-temporal datasets causes the model performance to decrease as we move away from the training set's sensing time. The performance of the model is bound to degrade as we move further away from the training period. Upon comparing the development of the slums in Kampala city using the time slider in Google Earth from the year 2014 to the dates at which satellite images are downloaded, change observations can be made. It can be seen that since 2014, the built density of slums has increased at pre-existing locations and smaller new patches have emerged at suburban areas of the city. Compared to the larger areas of already existing slums, these new settlements are quite smaller and hence could be ignored during prediction on the image tiles due to mixed classes with vegetation and formal areas.

To evaluate the results obtained, the prediction Accuracy obtained from the ***Baseline VGG3 + Data Augmentation*** CNN Model (90.3%) on the sentinel-2 imagery is a very good results based on the spatial resolution of the images. In [106], Federico .B and Damián .S also detect informal settlements using sentinel-2A image tiles. They employed the use of Remote Sensing Classifier to Predict tiles that belong to informal settlements. They recommend the use of deep learning techniques, like convolutional neural networks to improve the accuracy of the classification. This indicates that the CNN provides better classification accuracy especially

because it's an image based classifier. On another note, the use of building outline characteristics like size, shape and Nearest Neighbour distance provides for a lot more insight on the nature and morphology of an informal settlement.

Investigations on predicting informal settlements using road intersections from OSM data [20] showed that OSM data can be used in this scenario. However, some of the limitations to this study is that some informal settlements have roads that are not mapped on OSM which makes their detection difficult. This is where the use of Building feature characteristics, as shown in this research are useful. The characteristics help better delineate informal settlements. An even better alternative would be the combination of building feature characteristics and density of road intersections to identify these informal settlements. The use of Multilayer Perceptron (MLP) on hotspot analysis GiZscore and Nneighbours to predict informal settlements proved a good enough classifier, however in this research, though MLP performed well, Random Forest Classifier was a better classifier for Building feature characteristics to predict informality.

6.2. Limitations

While the study produced good research findings, there are several limitations and opportunities for further research Additional vector data of land parcels, small administrative district units, city´s functional zones, or land use data could be integrated to improve the creation of desired optimal shapes of image objects of compact homogeneous built-up areas for built-up density analysis. If available in accurate up-to-date form, vector layer of building footprints could be used for built-up density calculation, without the need of land cover classification and building extraction from the satellite image data. This layer would represent only buildings, without roads or other man-made urban structures. However, this data is not always freely available in accurate and up-to-date form, and this approach takes that into consideration. The lack of adequate data to test the model on other cities besides the two presented was experienced. This model worked for these two cities because there both that OSM data almost accurately available. This is not the case for most of the cities in the global south that are engulfed by informality.

7. CONCLUSION AND FUTURE WORKS

The main research Aim of this research work was to integrate OSM data and sentinel-2 imagery for classifying and monitoring the growth of informal settlements. To achieve this, the research questions were investigated:

1. How can we exploit the potential of using buildings outline characteristics such as size and shape to differentiate Informal settlements from Formal Settlements?

   The results from the hotspot clustering analysis demonstrated that informal settlements can be differentiated from formal areas based on the use of Building feature characteristics that is shape, size and NND. The informal settlement buildings are characterized by low values (cold spots) for each characteristic.

2. Is it possible to predict informal areas in a city by understanding housing informality in other cities of similar context using buildings outline characteristics and machine learning?

   Also its observed that though morphologically different, slums share some characteristics like the nature of building outlines. This is shown by the transferability of the classifier trained on one dataset to predict informal settlements in new datasets in the same city and in a different city.

3. What is the most appropriate Machine Learning technique based on accuracy to predict informal areas in a city based on buildings outline characteristics?

   The most appropriate ML technique is the Random Forest classifier. It proved its effectiveness in a broad range of applications of Geospatial technologies. As a non-parametric method, the key property of Random Forest is its capability to handle different statistical distributions of features, which was one of the main challenges of this study. This make Random Forest a suitable algorithm for our binary classification and prediction problem.

4. How can we exploit the potential of freely available Sentinel-2 satellite imagery with advanced machine learning to estimate the growth of Informal settlements?

Sentinel 2 images, despite having a spatial resolution of 10m, can be used for informal settlement mapping when coupled with Advanced Machine Learning Algorithms like CNNs.

We suggest that this research area could be explored more and future research could be done, building and improving on the results with emphasis on a better accurate representation of Informal settlements. Furthermore, other OSM features and parameters such as Roads and accessibility within the informal settlements or others could be used to better identify where the informality regions could be. This is because Informal settlements are usually characterized by narrow and short road segments this numerous dangles ant the end of the roads. Other parameters like proximity to hazardous areas like wetland and sewerage channels should be considered. This however, would however require much more detailed and complex land cover classification scheme as a source of information for this analysis, including identification of, vegetation types, difference between agricultural bare soil and bare soil areas in urban area or a construction sites and so on. If available, Digital Surface Model (DSM) could be used to more accurately identify and map informal settlements because inclusion of the third dimension of Elevation would help highlight more distinguishing characteristics of informal settlement buildings.

Nevertheless, the contribution of the temporal dimension to informal settlement mapping investigated in this research requires further testing with larger datasets and longer time series. With this mind, we think the positive results achieved in this research are worth building upon in future studies to further investigate the temporal domain as an input to informal settlement mapping and classification models.

BIBLIOGRAPHIC REFERENCES

1. Africa's Urban Transformation - Essays | Urban Age [Internet]. [cited 2020 Feb 6]. Available from: https://urbanage.lsecities.net/essays/africa-s-urban-transformation
2. 68% of the world population projected to live in urban areas by 2050, says UN | UN DESA | United Nations Department of Economic and Social Affairs [Internet]. [cited 2020 Feb 6]. Available from: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html
3. Dar es Salaam is growing rapidly—and its planners are struggling to keep up [Internet]. [cited 2020 Feb 6]. Available from: https://www.nationalgeographic.com/environment/2019/04/tanzanian-city-may-soon-be-one-of-the-worlds-most-populous/
4. About Slum Upgrading | Cities Alliance [Internet]. [cited 2020 Feb 6]. Available from: https://www.citiesalliance.org/about-slum-upgrading
5. Introduction to urban poverty | International Institute for Environment and Development [Internet]. [cited 2020 Feb 6]. Available from: https://www.iied.org/introduction-urban-poverty
6. Roy D, Lees MH, Palavalli B, Pfeffer K, Sloot MAP. The emergence of slums: A contemporary view on simulation models. Environ Model Softw [Internet]. 2014;59(2014):76–90. Available from: http://dx.doi.org/10.1016/j.envsoft.2014.05.004
7. File:Aerial view of Makoko Slum in Lagos Nigeria.png - Wikimedia Commons [Internet]. [cited 2020 Feb 6]. Available from: https://commons.wikimedia.org/wiki/File:Aerial_view_of_Makoko_Slum_in_Lagos_Nigeria.png
8. The 20 Worst Slums in Africa [Internet]. [cited 2020 Feb 6]. Available from: http://www.africaranking.com/20-worst-slums-in-africa/6
9. Mahabir R, Agouris P, Stefanidis A, Croitoru A, Crooks AT. Detecting and mapping slums using open data: a case study in Kenya. Int J Digit Earth [Internet]. 2018;0(0):1–25. Available from: https://doi.org/10.1080/17538947.2018.1554010
10. Suhartini N, Jones P. Urban Governance and Informal Settlements [Internet]. 2019. 235 p. Available from: https://www.springer.com/gp/book/9783030060930

11. Hofmann P, Taubenböck H, Werthmann C. Monitoring and modelling of informal settlements - A review on recent developments and challenges. 2015 Jt Urban Remote Sens Event, JURSE 2015. 2015;1–4.
12. Arimah BC. Slums as expression of social exclusion: Explaining the prevalence of slums in African countries. United Nations Hum Settlements Program [Internet]. 2001;1–33. Available from: http://www.oecd.org/dev/pgd/46837274.pdf
13. Kohli D, Sliuzas R, Stein A. Urban slum detection using texture and spatial metrics derived from satellite imagery. J Spat Sci [Internet]. 2016;61(2):405–26. Available from: http://dx.doi.org/10.1080/14498596.2016.1138247
14. Kohli D, Stein A, Sliuzas R. Uncertainty analysis for image interpretations of urban slums. Comput Environ Urban Syst [Internet]. 2016;60:37–49. Available from: http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.010
15. Wurm M, Taubenböck H, Weigand M, Schmitt A. Slum mapping in polarimetric SAR data using spatial features. Vol. 194, Remote Sensing of Environment. 2017. 190–204 p.
16. Stasolla M, Gamba P. Exploiting Spatial Patterns for Informal Settlement Detection in Arid Environments. Int Arch Photogramm Remote Sens Spat Inf Sci. 2007;36:31–6.
17. Ranguelova E, Weel B, Roy D, Kuffer M, Pfeffer K, Lees M. Image based classification of slums, built-up and non-built-up areas in Kalyan and Bangalore, India. Eur J Remote Sens [Internet]. 2019;52(sup1):40–61. Available from: https://doi.org/10.1080/22797254.2018.1535838
18. Detecting Informal Settlements using Satellite Imagery and Convolutional Neural Networks [Internet]. [cited 2020 Feb 6]. Available from: https://blog.goodaudience.com/detecting-informal-settlements-using-satellite-imagery-and-convolutional-neural-networks-d571a819bf44
19. Jean N, Xie M, Ermon S, Korea N. Combining Satellite Imagery and Machine Learning to Predict Poverty Inputs : Day # me satellite imagery Predic < ons : Economic indicators Learned features corresponding to buildings ,. 2015;4:2015.
20. Ibrahim MR, Titheridge H, Cheng T, Haworth J. predictSLUMS: A new model for identifying and predicting informal settlements and slums in cities from street intersections using machine learning. Comput Environ Urban Syst. 2019;76:31–56.
21. Amit Patel, Andrew Crooks and NK. Spatial Agent-based Modeling to Explore Slum

Formation Dynamics in Ahmedabad, India. 2018;37–54. Available from: http://link.springer.com/10.1007/978-3-319-59511-5

22. Crooks AT. Constructing and implementing an agent-based model of residential segregation through vector GIS. Int J Geogr Inf Sci. 2010;24(5):661–75.
23. United Nations Human Settlements Programme. The Challenge of Slums. The Challenge of Slums. 2012.
24. Ghani E. Urbanization and (In) Formalization. Urban Poverty Reduct Conf 2013 Bridg Rural Urban Perspect. 2013;
25. Bähr J, Mertins G. Marginalviertel in grossstadten der dritten welt. Geogr Rundsch. 2000;
26. The continuing challenge of informal settlements An introduction - Technische Informationsbibliothek (TIB) [Internet]. [cited 2020 Feb 6]. Available from: https://www.tib.eu/en/search/id/BLCP%3ACN063768801/The-continuing-challenge-of-informal-settlements/
27. Taubenböck H, Kraff NJ, Wurm M. The morphology of the Arrival City - A global categorization based on literature surveys and remotely sensed data. Appl Geogr. 2018;92(January):150–67.
28. Davis M. Planet of slums. New Left Review. 2004. 5–34 p.
29. Jamieson J. Shadow Cities: A Billion Squatters, a New Urban World. J Soc Polit Econ Stud. 2006;
30. Kohli D, Sliuzas R, Kerle N, Stein A. An ontology of slums for image-based classification. Comput Environ Urban Syst [Internet]. 2012;36(2):154–63. Available from: http://dx.doi.org/10.1016/j.compenvurbsys.2011.11.001
31. Kohli D. Identification and Characterization of Informal Settlements Using Satellite Images in Support of Land Administration.
32. Jokar Arsanjani J, Zipf A, Mooney P, Helbich M. OpenStreetMap in GIScience. OpenStreetMap GIScience Exp Res Appl. 2015;
33. Gervasoni L, Fenet S, Perrier R, Sturm P. Convolutional neural networks for disaggregated population mapping using open data. In: Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018. 2019.
34. Grippa T, Georganos S, Zarougui S, Bognounou P, Diboulo E, Forget Y, et al.

Mapping Urban Land Use at Street Block Level Using OpenStreetMap, Remote Sensing Data, and Spatial Metrics. ISPRS Int J Geo-Information. 2018;

35. Mahabir R, Croitoru A, Crooks A, Agouris P, Stefanidis A. News coverage, digital activism, and geographical saliency: A case study of refugee camps and volunteered geographical information. PLoS One. 2018;
36. Barrington-Leigh C, Millard-Ball A. The world's user-generated road map is more than 80% complete. PLoS One. 2017;
37. Fonte: Using openstreetmap to create land use and... - Google Scholar [Internet]. [cited 2020 Feb 6]. Available from: https://scholar.google.com/scholar_lookup?title=Using+Openstreetmap+to+Create+Land+Use+and+Land+Cover+Maps:+Development+of+an+Application&author=Fonte,+C.C.&author=Patriarca,+J.A.&author=Minghini,+M.&author=Antoniou,+V.&author=See,+L.&author=Brovelli,+M.A.&publication_year=2019&pages=1100–1123
38. OpenStreetMap: User-Generated Street Maps - IEEE Journals & Magazine [Internet]. [cited 2020 Feb 6]. Available from: https://ieeexplore.ieee.org/abstract/document/4653466
39. Johnson BA, Iizuka K. Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines. Appl Geogr [Internet]. 2016;67:140–9. Available from: http://dx.doi.org/10.1016/j.apgeog.2015.12.006
40. Yang D, Fu CS, Smith AC, Yu Q. Open land-use map: a regional land-use mapping strategy for incorporating OpenStreetMap with earth observations. Geo-Spatial Inf Sci [Internet]. 2017;20(3):269–81. Available from: http://doi.org/10.1080/10095020.2017.1371385
41. Kang L, Wang Q, Yan HW. Building extraction based on OpenStreetMap tags and very high spatial resolution image in urban area. Int Arch Photogramm Remote Sens Spat Inf Sci - ISPRS Arch. 2018;42(3):715–8.
42. Luo N, Wan T, Hao H, Lu Q. Fusing high-spatial-resolution remotely sensed imagery and OpenStreetMap data for land cover classification over urban areas. Remote Sens. 2019;11(1).
43. Huang H, Kieler B, Sester M. Urban building usage labeling by geometric and context analyses of the footprint data. Proceeding 26th Int Cartogr Conf. 2013;

44. Fan H, Zipf A, Fu Q. Estimation of building types on openstreetmap based on urban morphology analysis. In: Lecture Notes in Geoinformation and Cartography. 2014.
45. Henn A, Römer C, Gröger G, Plümer L. Automatic classification of building types in 3D city models. Geoinformatica. 2012;
46. Singleton AD, Spielman SE, Folch DC. Urban analytics. 176 p.
47. Roy D, Bernal D, Lees M. An exploratory factor analysis model for slum severity index in Mexico City. Urban Stud. 2019;
48. Weeks JR, Hill A, Stow D, Getis A, Fugate D. Can we spot a neighborhood from the air? Defining neighborhood structure in Accra, Ghana. In: GeoJournal. 2007.
49. Explore Our Data - Know Your City - SDI [Internet]. [cited 2020 Feb 6]. Available from: https://knowyourcity.info/explore-our-data/
50. Maula FK, Choerunnisa DN, Akbar R. Mapping Informal Settlements Using Geospatial Method. IOP Conf Ser Earth Environ Sci. 2019;385(1).
51. Gevaert CM, Persello C, Sliuzas R, Vosselman G. Informal settlement classification using point-cloud and image-based features from UAV data. ISPRS J Photogramm Remote Sens [Internet]. 2017;125:225–36. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2017.01.017
52. Gadiraju KK, Vatsavai RR, Kaza N, Wibbels E, Krishna A. Machine learning approaches for slum detection using very high resolution satellite images. IEEE Int Conf Data Min Work ICDMW. 2019;2018-Novem:1397–404.
53. Dovey K, Pafka E, Ristic M. Mapping Urbanities. Mapping Urbanities. 2017.
54. Gram-Hansen BJ, Azam F, Helber P, Coca-Castro A, Bilinski P, Varatharajan I, et al. Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data. AIES 2019 - Proc 2019 AAAI/ACM Conf AI, Ethics, Soc. 2019;361–8.
55. Mboga N, Persello C, Bergado JR, Stein A. Detection of informal settlements from VHR images using convolutional neural networks. Remote Sens. 2017;
56. Annemarie E, Carlos DW. Identification of appropriate data sources and analysis software to monitor the growth of informal settlements in Namibia. 2019 Jt Urban Remote Sens Event, JURSE 2019. 2019;1–4.
57. Kraff NJ, Taubenbock H, Wurm M. How dynamic are slums? EO-based assessment of Kibera's morphologic transformation. 2019 Jt Urban Remote Sens Event, JURSE

2019. 2019;6–9.

58. Hofmann P. Detecting Informal Settlements From Ikonos Image Data Using Methods Of Object Oriented Image Analysis – An Example From Cape Town (South Africa). Remote Sens Urban areas. 2001;
59. Hofmann P, Strobl J, Blaschke T, Kux H. Detecting informal settlements from QuickBird data in Rio de Janeiro using an object-based approach. Lect Notes Geoinf Cartogr. 2008;
60. Shekhar S. Detecting Slums From Quick Bird Data In Pune Using An Object Oriented Approach. ISPRS - Int Arch Photogramm Remote Sens Spat Inf Sci. 2012;
61. What is Hotspot Analysis? | Geospatiality [Internet]. [cited 2020 Feb 6]. Available from: https://glenbambrick.com/2016/01/21/what-is-hotspot-analysis/
62. Hot Spot Detection | Columbia University Mailman School of Public Health [Internet]. [cited 2020 Feb 16]. Available from: https://www.mailman.columbia.edu/research/population-health-methods/hot-spot-detection
63. Chakravorty S. Identifying crime clusters: The spatial principles. Middle States Geogr. 1995;
64. Naghibi SA, Pourghasemi HR, Dixon B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess. 2016;
65. Sadeghian A, Lim D, Karlsson J, Li J. Automatic target recognition using discrimination based on optimal transport. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2015.
66. Mollalo A, Sadeghian A, Israel GD, Rashidi P, Sofizadeh A, Glass GE. Machine learning approaches in GIS-based ecological modeling of the sand fly Phlebotomus papatasi, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran. Acta Trop. 2018;
67. Shirzadi MR, Mollalo A, Yaghoobi-Ershadi MR. Dynamic relations between incidence of Zoonotic Cutaneous Leishmaniasis and climatic factors in Golestan Province, Iran. J Arthropod Borne Dis. 2015;
68. Hatami M, Siahooei EA. Examines criteria applicable in the optimal location new cities, with approach for sustainable urban development. Middle East J Sci Res. 2013;

69. Bagheri M, Bazvand A, Ehteshami M. Application of artificial intelligence for the management of landfill leachate penetration into groundwater, and assessment of its environmental impacts. J Clean Prod. 2017;
70. Sadeghian A, Sundaram L, Wang DZ, Hamilton WF, Branting K, Pfeifer C. Automatic semantic edge labeling over legal citation graphs. Artif Intell Law. 2018;
71. Janalipour M, Mohammadzadeh A. A Fuzzy-GA Based Decision Making System for Detecting Damaged Buildings from High-Spatial Resolution Optical Images. Remote Sens. 2017;
72. Mohsen Shafieardekani, Mohsen Hatami. Forecasting Land Use Change in suburb by using Time series and Spatial Approach; Evidence from Intermediate Cities of Iran. Eur J Sci Res. 2013;
73. Which machine learning algorithm should I use? - The SAS Data Science Blog [Internet]. [cited 2020 Feb 6]. Available from: https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/
74. Crash Course in Convolutional Neural Networks for Machine Learning [Internet]. [cited 2020 Feb 15]. Available from: https://machinelearningmastery.com/crash-course-convolutional-neural-networks/
75. CS231n Convolutional Neural Networks for Visual Recognition [Internet]. [cited 2020 Feb 15]. Available from: http://cs231n.github.io/convolutional-networks/
76. Asmat A, Zamzami SZ. Automated House Detection and Delineation using Optical Remote Sensing Technology for Informal Human Settlement. Procedia - Soc Behav Sci. 2012;
77. Mahabir R, Croitoru A, Crooks A, Agouris P, Stefanidis A. A Critical Review of High and Very High-Resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities. Urban Sci. 2018;
78. Varshney KR, Chen GH, Abelson B, Nowocin K, Sakhrani V, Xu L, et al. Targeting Villages for Rural Development Using Satellite Image Analysis. Big Data. 2015;
79. Xie M, Jean N, Burke M, Lobell D, Ermon S. Transfer learning from deep features for remote sensing and poverty mapping. In: 30th AAAI Conference on Artificial Intelligence, AAAI 2016. 2016.
80. Gevaert CM, Persello C, Sliuzas R, Vosselman G. Classification Of Informal

Settlements Through The Integration Of 2d And 3d Features Extracted From Uav Data. ISPRS Ann Photogramm Remote Sens Spat Inf Sci. 2016;

81. Kuffer M, Pfeffer K, Sliuzas R. Slums from space-15 years of slum mapping using remote sensing. Remote Sensing. 2016.
82. Stasolla M, Gamba P. Spatial indexes for the extraction of formal and informal human settlements from high-resolution SAR images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2008;
83. Chen L, Xie T, Wang X, Wang C. Identifying urban villages from city-wide satellite imagery leveraging mask r-Cnn. UbiComp/ISWC 2019- - Adjun Proc 2019 ACM Int Jt Conf Pervasive Ubiquitous Comput Proc 2019 ACM Int Symp Wearable Comput. 2019;29–32.
84. (No Title) [Internet]. [cited 2020 Feb 6]. Available from: https://www.ubos.org/wp-content/uploads/publications/2014CensusProfiles/KAMPALA-KCCA.pdf
85. Dar Es Salaam Population 2020 (Demographics, Maps, Graphs) [Internet]. [cited 2020 Feb 6]. Available from: http://worldpopulationreview.com/world-cities/dar-es-salaam-population/
86. Kuffer M, Persello C, Pfeffer Itc K, Sliuzas R, Rao V. Do we underestimate the global slum population? 2019 Jt Urban Remote Sens Event, JURSE 2019. 2019;1–4.
87. UN-Habitat. Implementing the New Urban Agenda. 2016;48. Available from: https://unhabitat.org/books/21-projects-compendium-implementing-the-new-urban-agenda/#
88. GEOFABRIK // Home [Internet]. [cited 2020 Feb 6]. Available from: http://www.geofabrik.de/
89. GitHub - microsoft/CNTK: Microsoft Cognitive Toolkit (CNTK), an open source deep-learning toolkit [Internet]. [cited 2020 Feb 6]. Available from: https://github.com/Microsoft/CNTK
90. Microsoft releases 18M building footprints in Uganda and Tanzania to enable AI Assisted Mapping | Maps Blog [Internet]. [cited 2020 Feb 6]. Available from: https://blogs.bing.com/maps/2019-09/microsoft-releases-18M-building-footprints-in-uganda-and-tanzania-to-enable-ai-assisted-mapping
91. OVERKILL (Command) | AutoCAD LT 2019 | Autodesk Knowledge Network [Internet]. [cited 2020 Feb 6]. Available from:

https://knowledge.autodesk.com/support/autocad-lt/learn-explore/caas/CloudHelp/cloudhelp/2019/ENU/AutoCAD-LT/files/GUID-44B9ECFC-752C-4CC5-9DA3-84DBF3B17CA6-htm.html

92. Pesaresi M, Corbane C, Julea A, Florczyk AJ, Syrris V, Soille P. Assessment of the added-value of sentinel-2 for detecting built-up areas. Remote Sens. 2016;
93. Verma D, Jana A, Ramamritham K. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. Habitat Int. 2019;
94. Gómez JA, Patiño JE, Duque JC, Passos S. Spatiotemporal Modeling of Urban Growth Using Machine Learning. Remote Sens. 2019;12(1):109.
95. Kuc G, Chormański J. Sentinel-2 Imagery for Mapping and Monitoring Imperviousness in Urban Areas. ISPRS - Int Arch Photogramm Remote Sens Spat Inf Sci. 2019;XLII-1/W2:43–7.
96. ESA Standard Document | Enhanced Reader [Internet]. [cited 2020 Feb 16]. Available from: moz-extension://e1f079c0-3081-424a-a3d2-ff14239a84ce/enhanced-reader.html?openApp&pdf=https%3A%2F%2Fsentinel.esa.int%2Fdocuments%2F247904%2F685211%2FSentinel-2_User_Handbook
97. Richmond A, Myers I, Namuli H. Urban Informality and Vulnerability: A Case Study in Kampala, Uganda. Urban Sci. 2018;2(1):22.
98. Jana B, Mohanty SN. An Intuitionistic Fuzzy Logic Models for Multicriteria Decision Making Under Uncertainty. J Inst Eng Ser C. 2017;98(2):197–201.
99. Yan X, Ai T, Yang M, Yin H. A graph convolutional neural network for classification of building patterns using spatial vector data. ISPRS J Photogramm Remote Sens [Internet]. 2019;150(February):259–73. Available from: https://doi.org/10.1016/j.isprsjprs.2019.02.010
100. Taubenböck H, Kraff NJ. The physical face of slums: A structural comparison of slums in Mumbai, India, based on remotely sensed data. J Hous Built Environ. 2014;29(1):15–38.
101. Optimized Hot Spot Analysis—ArcGIS Pro | Documentation [Internet]. [cited 2020 Feb 6]. Available from: https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/optimized-hot-spot-analysis.htm
102. Independent Samples t Test - SPSS Tutorials - LibGuides at Kent State University [Internet]. [cited 2020 Feb 15]. Available from:

https://libguides.library.kent.edu/SPSS/IndependentTTest

103. Boots B. Developing local measures of spatial association for categorical data. J Geogr Syst. 2003;

104. On solid ground: armed with land titles, Tanzania's slum dwellers tackle poverty - The East African [Internet]. [cited 2020 Feb 17]. Available from: https://www.theeastafrican.co.ke/news/ea/Tanzania-slum-dwellers-tackle-poverty/4552908-4969252-nsc3esz/index.html

105. Tanzania - Population Living In Slums (% Of Urban Population) - 1990-2014 Data | 2020 Forecast [Internet]. [cited 2020 Feb 17]. Available from: https://tradingeconomics.com/tanzania/population-living-in-slums-percent-of-urban-population-wb-data.html

106. Detecting informal settlements in South America: How I built it [Internet]. [cited 2020 Feb 20]. Available from: https://blog.mapbox.com/detecting-informal-settlements-in-south-america-how-i-built-it-cb139a870816

8. ANNEXES

8.1. Independent Samples T-Tests On Hotspot Analysis GiZScores and Nneighbors

SIZE(AREA)

Group Statistics

Group		N	Mean	Std. Deviation	Std. Error Mean
GiZScore	ColdSpot(Informal)	104884	-4.4855	2.0227	0.0062
	HotSpot (Formal)	36868	3.6859	1.8595	0.0097
Nneighbors	ColdSpot(Informal)	104884	958.00	302.407	.934
	HotSpot (Formal)	36868	363.13	129.434	.674

Independent Samples Test

		Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
GiZScore	Equal variances assumed	1175.325	.000	-681.104	141750	0.000	-8.1714	0.0120	-8.1949	-8.1479
	Equal variances not assumed			-709.098	69673.507	0.000	-8.1714	0.0115	-8.1940	-8.1488
Nneighbors	Equal variances assumed	12231.725	0.000	366.098	141750	0.000	594.864	1.625	591.679	598.049
	Equal variances not assumed			516.526	136905.311	0.000	594.864	1.152	592.607	597.121

NEAREST NEIGHBOUR DISTANCE (NND)

Group Statistics

Group		N	Mean	Std. Deviation	Std. Error Mean
GiZScore	ColdSpot(Informal)	123467	-5.6911	2.8813	0.0082
	HotSpot (Formal)	41942	5.1848	3.3074	0.0161
Nneighbors	ColdSpot(Informal)	123467	936.37	281.848	.802
	HotSpot (Formal)	41942	344.97	119.569	.584

Independent Samples Test

		Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
GiZScore	Equal variances assumed	92.312	.000	-642.507	165407	0.000	-10.8760	0.0169	-10.9092	-10.8428
	Equal variances not assumed			-600.476	64888.994	0.000	-10.8760	0.0181	-10.9115	-10.8405
Nneighbors	Equal variances assumed	14527.675	0.000	417.164	165407	0.000	591.403	1.418	588.625	594.182
	Equal variances not assumed			596.111	158215.047	0.000	591.403	.992	589.459	593.348

SHAPE (VERTEX NUMBER)

Group Statistics

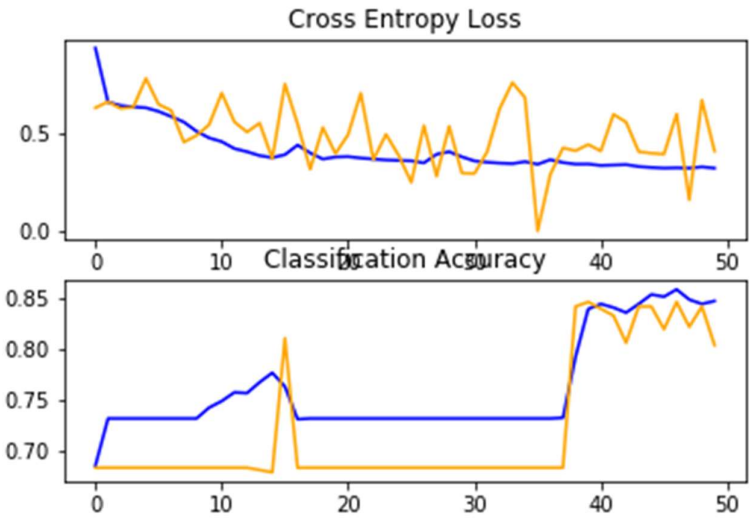
Group		N	Mean	Std. Deviation	Std. Error Mean
GiZScore	ColdSpot(Informal)	129377	-4.8803	1.9626	0.0055
	HotSpot (Formal)	34552	5.8729	4.6398	0.0250
Nneighbors	ColdSpot(Informal)	129377	850.26	319.712	.889
	HotSpot (Formal)	34552	512.44	286.337	1.540

Independent Samples Test

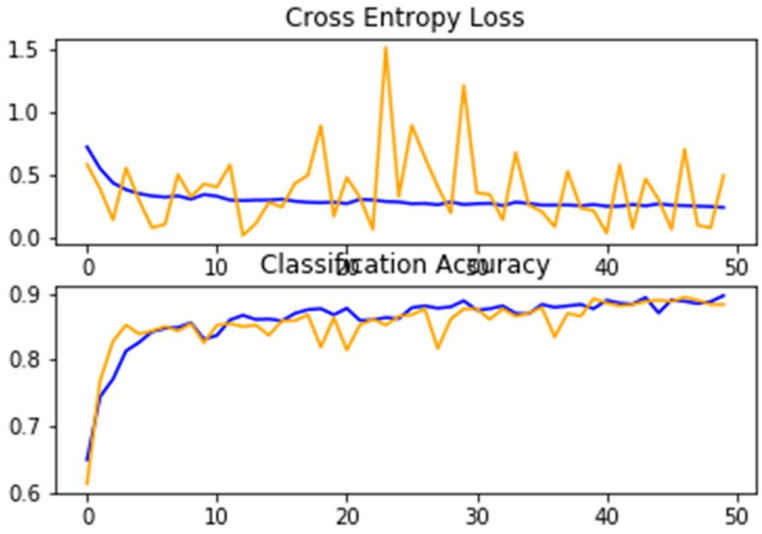
		Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
GiZScore	Equal variances assumed	20595.280	0.000	-645.086	163927	0.000	-10.7532	0.0167	-10.7859	-10.7206
	Equal variances not assumed			-420.864	37908.754	0.000	-10.7532	0.0256	-10.8033	-10.7031
Nneighbors	Equal variances assumed	66.235	.000	178.245	163927	0.000	337.822	1.895	334.107	341.537
	Equal variances not assumed			189.950	59623.581	0.000	337.822	1.778	334.336	341.308

8.2. Plots for Accuracy and Loss for Some CNN Models - Train (blue) and Test (orange) datasets.

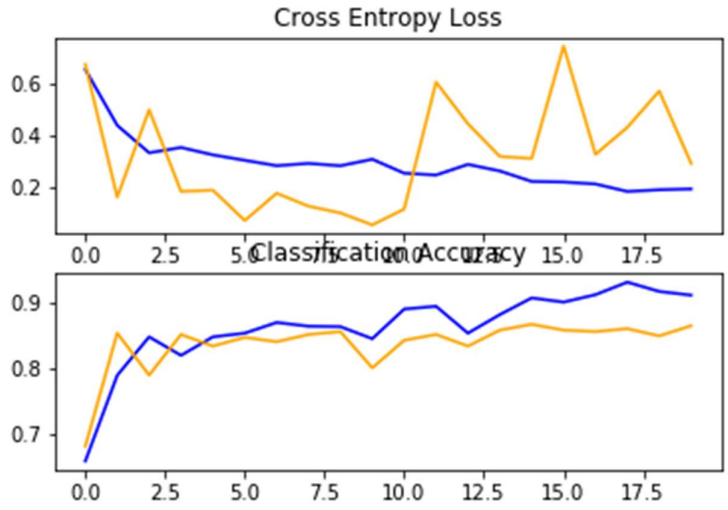
8.2.1. 1-VGG



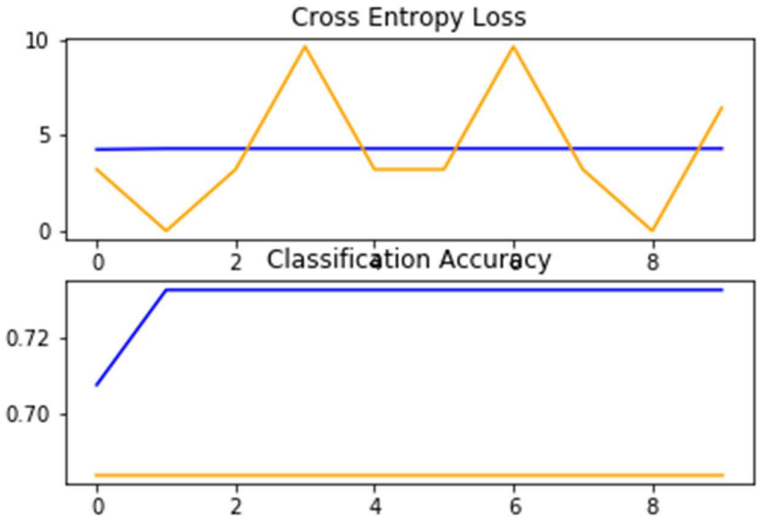
8.2.2. 3-VGG



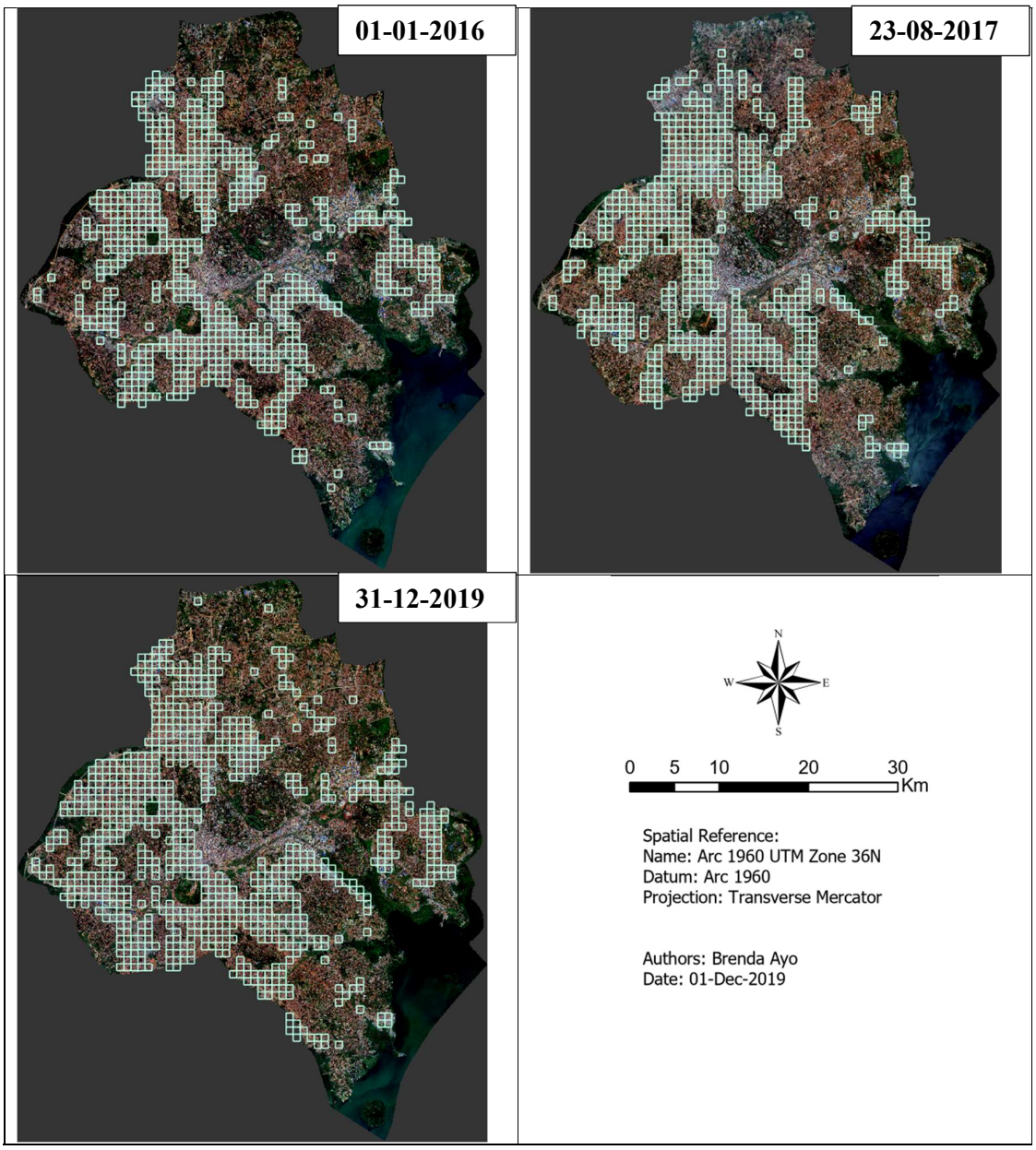
8.2.3. Baseline VGG3 + Data Augmentation



8.2.4. Pre-Trained VGG16



8.3. Informality Regions for Kampala in 2016 (top left), 2017 (top right) and 2019 (bottom)



Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS